A Study on
Multimodal and Interactive Explanations
for Visual Question Answering

Kamran Alipour, Jurgen P. Schulze, Yi Yao, Avi Ziskind, and Giedrius Burachas

1UC San Diego
2SRI International
Introduction

Visual Question Answering

- Answering a question in natural language regarding a given image
- Attention-based models on image features influenced by the question to produce answers

What is the animal in this picture?

Zebra.
Introduction

eXplainable AI (XAI)

○ Used for years in medicine, robotics, ...

○ Mainly attention-based

○ Various formats: visual, textual, ...
Contributions

- Generated multiple explanation modes from attention features and annotations
- Introduced an interactive explanation mode
- Prediction task: A novel assessment of the efficacy of explanations
- A user study to evaluate our XAI system
eXplainable Visual Question Answering (XVQA)

What is in the dog's mouth? 

ResNet -> LSTM -> Question Features -> Attention Layer -> FC -> Answer

Image Features

Frisbee
User Study

**Input**

- Images and Questions
- Explanations
- VQA correctness and confidence

**TASK:**

- Rank explanations
- Predict if VQA answers correctly based on explanations.
User Study

<table>
<thead>
<tr>
<th>Subjects:</th>
<th>90</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trials:</td>
<td>10513</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Group</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NE</td>
<td>Control group</td>
</tr>
<tr>
<td>SP</td>
<td>Spatial attention</td>
</tr>
<tr>
<td>SA</td>
<td>Active (steerable) attention</td>
</tr>
<tr>
<td>SE</td>
<td>Semantic</td>
</tr>
<tr>
<td>OA</td>
<td>Object attention</td>
</tr>
<tr>
<td>AL</td>
<td>All explanations</td>
</tr>
</tbody>
</table>
Explanation modes

Spatial attention

Shows the parts of the image the model is focused on while preparing the answer.

**QUESTION:** Where is this place?

**ANSWER:** Airport.
Explanation modes

Active (streerable) attention

What covers the ground?

ResNet

LSTM

Question Features

Image Features

Attention Layer

Multiplied to image features

Answer

Snow

Attention map drawn on the image by the user
Explanation modes

Semantic:

Bounding box

Scene graph
Explanation modes

Semantic:

Textual Explanation [Ghosh et al. 2019]

QUESTION: What is this sport?

ANSWER: Soccer.

Explanations:
Because the image contains: white line on soccer field
Because the image contains: man on soccer field
Because the image contains: soccer shorts with numbers
Explanation modes

Object attention [Ray et al. 2019]

**QUESTION:** What food is he eating?

**ANSWER:** Sandwich.
Results

Prediction accuracies in different study groups
Results

Prediction accuracy vs. explanation ratings when system accurate
Results

Prediction accuracy vs. explanation ratings when system inaccurate
Results

User Prediction Accuracy (\%) vs Trials

- Exp. Blocks
- Control Blocks

Prediction accuracy progress
Results

User confidence progress: Active attention vs. other explanations
Discussion

- Interactive explanations improved users confidence compared to other explanations
- When AI is inaccurate, explanations are more helpful
- Getting exposed to multiple explanation modes can be conflicting/overwhelming and reduce prediction accuracy
Conclusion and Future Work

- Interactive experiment to probe explanation effectiveness
- Explanations help accuracy prediction
- Users confidence improved when exposed to explanations
Thank You!
## User Study Statistics

<table>
<thead>
<tr>
<th>Group</th>
<th>Subjects</th>
<th>Trials</th>
</tr>
</thead>
<tbody>
<tr>
<td>NE</td>
<td>Control group</td>
<td>15</td>
</tr>
<tr>
<td>SP</td>
<td>Spatial attention</td>
<td>15</td>
</tr>
<tr>
<td>SA</td>
<td>Active (steerable) attention</td>
<td>15</td>
</tr>
<tr>
<td>SE</td>
<td>Semantic</td>
<td>15</td>
</tr>
<tr>
<td>OA</td>
<td>Object attention</td>
<td>15</td>
</tr>
<tr>
<td>AL</td>
<td>All explanations</td>
<td>15</td>
</tr>
<tr>
<td><strong>Total</strong>:</td>
<td></td>
<td><strong>90</strong></td>
</tr>
</tbody>
</table>
User Study Structure

Control Group
- Practice Block
  - No-Exp Trial
  - No-Exp Trial
- Control Block
  - No-Exp Trial
  - No-Exp Trial
  - No-Exp Trial
  - No-Exp Trial
  - No-Exp Trial

Explanation Groups
- Practice Block
  - Exp Trial
  - No-Exp Trial
- Explanation Block
  - 5 x Exp Trial
- Control Block
  - 5 x No-Exp Trial
- Explanation Block
  - 5 x Exp Trial
- Control Block
  - 5 x No-Exp Trial
User Study Structure

Input

- Images and Questions
- Explanations
- VQA correctness and confidence

TASK:

- Rank explanations
- Predict if VQA answers correctly based on explanations.