# What’s in a Question: Using Visual Questions as a Form of Supervision

Siddha Ganju, Olga Russakovsky, Abhinav Gupta

## Questions are informative

<table>
<thead>
<tr>
<th>What breed of dog is this?</th>
<th>This question suggests:</th>
</tr>
</thead>
<tbody>
<tr>
<td>- the animal in the scene is a dog</td>
<td></td>
</tr>
<tr>
<td>- breed is a property of dog</td>
<td></td>
</tr>
<tr>
<td>- all dogs in the scene are the same breed</td>
<td></td>
</tr>
<tr>
<td>- knowing the breed may be important</td>
<td></td>
</tr>
</tbody>
</table>

**Goal:** quantify and utilize this information

### Questions → image captions

**Questions:**
- Was this picture taken during the day?
- What are these two people doing?
- What color is the right person’s hat?

**Generated caption:** people during day with hat

### Questions → object classification

**Inferred objects:**
- Fire Hydrant
- What is in the water?
- Fish
- Plastic bag

**Questions (on COCO):** 29.3% recall, 82.4% precision

## Method for improving VQA

### Training data

**Q:** What is under the plane?

**A:** Water

**Other questions:**
- Can this plane land on water?
- How many planes are there?

### Model: iBOWIMG-2x

**CNN** → **Text Embedding** → **Multiple choice**
- dolphin (0.2)
- yes (0.1)
- water (0.7)

**Can this plane land on water?**

**How many planes are there?**

### Using other questions in iBOWIMG-2x

**Training:**
- What breed of dog is this?
- What is under the plane, Can this plane… are there?

**Training augmented:**
- What breed of dog is this?
- What is under the plane, Can this plane… are there?
- What is under the plane, How many planes are there?
- What is under the plane, (0.2)

**Testing without:**
- Is this cat lying on a sofa, (0.2)

**Testing with:**
- Is this cat lying on a sofa, What color is the car seat?

## Results on VQA

### VQA dataset v1.0, 3 answered questions per training image, 3 questions per test image

#### Experiment #1: with unanswered questions

<table>
<thead>
<tr>
<th>Model</th>
<th>Training?</th>
<th>Test?</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>iBOWIMG-2x</td>
<td>yes</td>
<td>—</td>
<td>47.3</td>
</tr>
<tr>
<td>iBOWIMG-2x, augmented</td>
<td>yes</td>
<td>—</td>
<td>50.4</td>
</tr>
<tr>
<td>iBOWIMG-2x</td>
<td>yes</td>
<td>yes</td>
<td>50.9</td>
</tr>
</tbody>
</table>

### Experiment #2: standard benchmark

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Yes/no Accuracy</th>
<th>Number Accuracy</th>
<th>Word Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>iBOWIMG-2x</td>
<td>62.8</td>
<td>80.7</td>
<td>37.9</td>
<td>53.1</td>
</tr>
</tbody>
</table>

## Questions → image captions

### Input → Model → Output

<table>
<thead>
<tr>
<th>3 questions</th>
<th>Seq2seq [2]</th>
<th>Image caption</th>
<th>0.140</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 question</td>
<td>Used directly</td>
<td>Image caption</td>
<td>0.058</td>
</tr>
<tr>
<td>Image</td>
<td>Neural Talk [3]</td>
<td>Image caption</td>
<td>0.194</td>
</tr>
</tbody>
</table>

## Bibliography