Ask Me Anything: Free-form Visual Question Answering
Based on Knowledge from External Sources
Supplementary Material

Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, Anton van den Hengel
School of Computer Science, The University of Adelaide, Australia
{qi.wu01,p.wang,chunhua.shen,anthony.dick,anton.vandenhengel}@adelaide.edu.au

In this supplementary material, we provide additional VQA examples generated using our system. The questions in the table 1, 2 and table 3 all start with ‘why ...?’. To answer these questions would typically require human common-sense or even high-level knowledge. We display a set of examples for which our final model gives the right answer while the base line model VggNet-LSTM generates the wrong answer. Table 4 and table 5 shows examples where our final model gives the right answer while the base line model VggNet-LSTM generates the wrong answer, for various question types. Table 6 provides some failure examples produced by our system.

<table>
<thead>
<tr>
<th>Why do they have umbrellas?</th>
<th>Why is he swinging backhand?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ours:</strong> raining</td>
<td><strong>Ours:</strong> raining</td>
</tr>
<tr>
<td><strong>Vgg+LSTM:</strong> yes</td>
<td><strong>Vgg+LSTM:</strong> yes</td>
</tr>
<tr>
<td><strong>Ground Truth:</strong> raining</td>
<td><strong>Ground Truth:</strong> raining</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Why are they wearing such bright colors?</th>
<th>Why are the men wearing orange?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ours:</strong> safety</td>
<td><strong>Ours:</strong> safety</td>
</tr>
<tr>
<td><strong>Vgg+LSTM:</strong> yes</td>
<td><strong>Vgg+LSTM:</strong> yes</td>
</tr>
<tr>
<td><strong>Ground Truth:</strong> safety</td>
<td><strong>Ground Truth:</strong> safety</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Why is the man jumping?</th>
<th>Why is this room warm?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ours:</strong> skateboarding</td>
<td><strong>Ours:</strong> fireplace</td>
</tr>
<tr>
<td><strong>Vgg+LSTM:</strong> unknown</td>
<td><strong>Vgg+LSTM:</strong> to sleep</td>
</tr>
<tr>
<td><strong>Ground Truth:</strong> skateboarding</td>
<td><strong>Ground Truth:</strong> fireplace</td>
</tr>
</tbody>
</table>

Table 1. Some examples for which our final model gives the right answer while the base line model VggNet-LSTM generates the wrong answer. All questions start with ‘why’ and some can only be answered with common sense knowledge.
<table>
<thead>
<tr>
<th>Question</th>
<th>Ours</th>
<th>Vgg+LSTM</th>
<th>Ground Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Why is this person wet?</td>
<td>surfing</td>
<td>beach</td>
<td>surfing</td>
</tr>
<tr>
<td>Why is the baby wearing a snowsuit?</td>
<td>cold</td>
<td>safety</td>
<td>cold</td>
</tr>
<tr>
<td>Why does the boy have his arms in that position?</td>
<td>balance</td>
<td>to catch</td>
<td>balance</td>
</tr>
<tr>
<td>Why are two of the giraffes so much shorter than the other three?</td>
<td>they are babies</td>
<td>yes</td>
<td>babies</td>
</tr>
<tr>
<td>Why do these sheep have paint on them?</td>
<td>identification</td>
<td>to eat</td>
<td>identification</td>
</tr>
<tr>
<td>Why is this ground white?</td>
<td>snow</td>
<td>cold</td>
<td>snow</td>
</tr>
<tr>
<td>Why is there sand around the orange object?</td>
<td>safety</td>
<td>to balance</td>
<td>safety</td>
</tr>
<tr>
<td>Why is the man wearing black there?</td>
<td>umpire</td>
<td>safety</td>
<td>umpire</td>
</tr>
<tr>
<td>Why is she wearing a potholder on her arm?</td>
<td>cooking</td>
<td>train</td>
<td>cooking</td>
</tr>
<tr>
<td>Why is the road closed?</td>
<td>train</td>
<td>winter</td>
<td>train</td>
</tr>
<tr>
<td>Why are there no leaves on the trees?</td>
<td>winter</td>
<td>unknown</td>
<td>winter</td>
</tr>
<tr>
<td>Why does he have glasses on?</td>
<td>to see</td>
<td>to be</td>
<td>to see</td>
</tr>
<tr>
<td>Why is the ground wet?</td>
<td>rain</td>
<td>resting</td>
<td>rain</td>
</tr>
<tr>
<td>Why are the animals laying here?</td>
<td>resting</td>
<td>playing frisbee</td>
<td>resting</td>
</tr>
<tr>
<td>Why is the man running?</td>
<td>playing frisbee</td>
<td>running</td>
<td>running</td>
</tr>
<tr>
<td>Why is the cat sitting on the bench?</td>
<td>resting</td>
<td>to sleep</td>
<td>resting</td>
</tr>
<tr>
<td>Why are her hands in the air?</td>
<td>flying kite</td>
<td>playing tennis</td>
<td>flying kite</td>
</tr>
<tr>
<td>Why is the man standing?</td>
<td>surfing</td>
<td>tennis ball</td>
<td>to eat</td>
</tr>
<tr>
<td>Why is the child running?</td>
<td>flying kite</td>
<td>playing frisbee</td>
<td>to eat</td>
</tr>
<tr>
<td>Why is there a giraffe in this setting?</td>
<td>flying kite</td>
<td>playing tennis</td>
<td>zoo</td>
</tr>
</tbody>
</table>

Table 3. Some examples for which our final model gives the right answer while the base line model VggNet-LSTM generates the wrong answer. All questions start with ‘why’ and some can only can be answered with common sense knowledge.
<table>
<thead>
<tr>
<th>Question</th>
<th>Ours</th>
<th>Vgg+LSTM</th>
<th>Ground Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>What kind of weather it is?</td>
<td>sunny</td>
<td>cloudy</td>
<td>sunny</td>
</tr>
<tr>
<td>What is in the cup?</td>
<td>coffee</td>
<td>wine</td>
<td>coffee</td>
</tr>
<tr>
<td>What kind of room is this?</td>
<td>bedroom</td>
<td>living</td>
<td>bedroom</td>
</tr>
<tr>
<td>Where did the water come from?</td>
<td>ocean</td>
<td>beach</td>
<td>ocean</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Question</th>
<th>Ours</th>
<th>Vgg+LSTM</th>
<th>Ground Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>What game is being played on the beach?</td>
<td>volleyball</td>
<td>soccer</td>
<td>volleyball</td>
</tr>
<tr>
<td>How many busses are there?</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Is the person wearing a shirt?</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>What is the colorful object in the middle of the image?</td>
<td>kite</td>
<td>frisbee</td>
<td>kite</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Question</th>
<th>Ours</th>
<th>Vgg+LSTM</th>
<th>Ground Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>What are the children holding?</td>
<td>teddy bears</td>
<td>wii</td>
<td>teddy bears</td>
</tr>
<tr>
<td>Where is this picture?</td>
<td>market</td>
<td>on left</td>
<td>market</td>
</tr>
<tr>
<td>What kind of cheese is this?</td>
<td>mozzarella</td>
<td>chicken</td>
<td>mozzarella</td>
</tr>
<tr>
<td>What room is this?</td>
<td>bathroom</td>
<td>kitchen</td>
<td>bathroom</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Question</th>
<th>Ours</th>
<th>Vgg+LSTM</th>
<th>Ground Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>What style of cooking is this?</td>
<td>chinese</td>
<td>pizza</td>
<td>chinese</td>
</tr>
<tr>
<td>Is this a men’s room or a women’s room?</td>
<td>men’s</td>
<td>hotel</td>
<td>men’s</td>
</tr>
<tr>
<td>What is on the top of the animals’ heads?</td>
<td>horns</td>
<td>rocks</td>
<td>horns</td>
</tr>
<tr>
<td>Is this a vegetable?</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Question</th>
<th>Ours</th>
<th>Vgg+LSTM</th>
<th>Ground Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>What are the cats sleeping on?</td>
<td>car</td>
<td>table</td>
<td>car</td>
</tr>
<tr>
<td>Is it safe for the pedestrians to cross the street?</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>What other word is written for this sign?</td>
<td>stop</td>
<td>new</td>
<td>stop</td>
</tr>
<tr>
<td>How many airplanes?</td>
<td>4</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 4. Some examples for which our final model gives the right answer while the base line model VggNet-LSTM generates the wrong answer. Various question types are shown.
<table>
<thead>
<tr>
<th>What is he looking at?</th>
<th>What kind of meat is on this?</th>
<th>Which game is being played?</th>
<th>What brand is the bat bag?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ours:</strong> toothbrush</td>
<td>bacon</td>
<td>soccer</td>
<td>nike</td>
</tr>
<tr>
<td><strong>Vgg+LSTM:</strong> camera</td>
<td>chicken</td>
<td>tennis</td>
<td>wilson</td>
</tr>
<tr>
<td><strong>Ground Truth:</strong> toothbrush</td>
<td>bacon</td>
<td>soccer</td>
<td>nike</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Is this inside?</th>
<th>What season does it look like?</th>
<th>Is this a healthy breakfast?</th>
<th>Is this meal healthy?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ours:</strong> no</td>
<td>fall</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td><strong>Vgg+LSTM:</strong> yes</td>
<td>winter</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td><strong>Ground Truth:</strong> no</td>
<td>fall</td>
<td>no</td>
<td>yes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>The green item on the pizza, what is it called?</th>
<th>Does this animal have fur?</th>
<th>Is this a home office?</th>
<th>What letter is inside the blue circle?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ours:</strong> broccoli</td>
<td>no</td>
<td>yes</td>
<td>p</td>
</tr>
<tr>
<td><strong>Vgg+LSTM:</strong> carrots</td>
<td>yes</td>
<td>no</td>
<td>b</td>
</tr>
<tr>
<td><strong>Ground Truth:</strong> broccoli</td>
<td>no</td>
<td>yes</td>
<td>p</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>What type of food is this person eating?</th>
<th>In what type of establishment is this taken?</th>
<th>Is that meat on the plate?</th>
<th>What is being celebrated?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ours:</strong> donut</td>
<td>zoo</td>
<td>yes</td>
<td>birthday</td>
</tr>
<tr>
<td><strong>Vgg+LSTM:</strong> pizza</td>
<td>zebra</td>
<td>no</td>
<td>pizza</td>
</tr>
<tr>
<td><strong>Ground Truth:</strong> donut</td>
<td>zoo</td>
<td>yes</td>
<td>birthday</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>What kind of building is this?</th>
<th>Is the weather cold or warm?</th>
<th>Is that normal a banana on a record?</th>
<th>What color is the snow?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ours:</strong> barn</td>
<td>cold</td>
<td>no</td>
<td>white</td>
</tr>
<tr>
<td><strong>Vgg+LSTM:</strong> church</td>
<td>sunny</td>
<td>yes</td>
<td>blue</td>
</tr>
<tr>
<td><strong>Ground Truth:</strong> barn</td>
<td>cold</td>
<td>no</td>
<td>white</td>
</tr>
</tbody>
</table>

Table 5. Some examples for which our final model gives the right answer while the base line model VggNet-LSTM generates the wrong answer. Various question types are shown.
<table>
<thead>
<tr>
<th>What season is it?</th>
<th>What shoe company is advertised?</th>
<th>Is this guy going to jump high?</th>
<th>What is on the keyboard?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours: fall</td>
<td>vans</td>
<td>no</td>
<td>mouse</td>
</tr>
<tr>
<td>Ground Truth:</td>
<td>winter</td>
<td>yes</td>
<td>cat</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>What color is the front of the tow truck?</th>
<th>What kind of wood is the table made of?</th>
<th>Where is the telephone?</th>
<th>How deep is water?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours: red</td>
<td>oak</td>
<td>on desk</td>
<td>shallow</td>
</tr>
<tr>
<td>Ground Truth: white</td>
<td>cherry</td>
<td>on nightstand</td>
<td>10 feet</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Who ate some of the cake?</th>
<th>What city is this?</th>
<th>What utensils are shown?</th>
<th>What is the building facade made from?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours: man</td>
<td>new york</td>
<td>fork and knife</td>
<td>brick</td>
</tr>
<tr>
<td>Ground Truth: person</td>
<td>las vegas</td>
<td>fork</td>
<td>stone</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>What is she holding in her hand?</th>
<th>What creature is this?</th>
<th>What are the colors of the court?</th>
<th>What is on their hand?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours: ski poles</td>
<td>horse</td>
<td>blue</td>
<td>hot dog</td>
</tr>
<tr>
<td>Ground Truth: ski pole</td>
<td>pegasus</td>
<td>blue and green</td>
<td>glove</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>What's on the floor?</th>
<th>Who took this photo?</th>
<th>What food is this, really?</th>
<th>What kind of weather is this?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours: scissors</td>
<td>photographer</td>
<td>chicken</td>
<td>rainy</td>
</tr>
<tr>
<td>Ground Truth: tape</td>
<td>christopher brown</td>
<td>cake</td>
<td>cloudy</td>
</tr>
</tbody>
</table>