

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.





	
Why is she wearing a crown?	Why is he smiling?
<i>Ours:</i> birthday	<i>Ours:</i> happy
<i>Vgg+LSTM:</i> to eat	<i>Vgg+LSTM:</i> unknown
<i>Ground Truth:</i> birthday	<i>Ground Truth:</i> happy
	
Why is the zebra on the ground?	Why is a man sitting under an umbrella?
<i>Ours:</i> resting	<i>Ours:</i> shade
<i>Vgg+LSTM:</i> eat	<i>Vgg+LSTM:</i> safety
<i>Ground Truth:</i> resting	<i>Ground Truth:</i> shade

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.

	
Why do they have umbrellas	Why is he swinging backhand?
<i>Ours:</i> raining	<i>Ours:</i> to hit ball
<i>Vgg+LSTM:</i> yes	<i>Vgg+LSTM:</i> tennis ball
<i>Ground Truth:</i> raining	<i>Ground Truth:</i> to hit ball
	
Why are they wearing such bright colors?	Why are the men wearing orange?
<i>Ours:</i> safety	<i>Ours:</i> team
<i>Vgg+LSTM:</i> yes	<i>Vgg+LSTM:</i> to
<i>Ground Truth:</i> safety	<i>Ground Truth:</i> team
	
Why is the man jumping?	Why is this room warm?
<i>Ours:</i> skateboarding	<i>Ours:</i> fireplace
<i>Vgg+LSTM:</i> unknown	<i>Vgg+LSTM:</i> to sleep
<i>Ground Truth:</i> skateboarding	<i>Ground Truth:</i> fireplace

Table 2. 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 of them can only be answered with common sense knowledge.











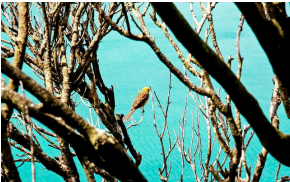











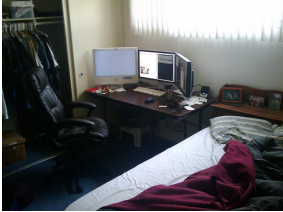











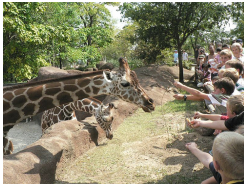

			
Why is this person wet?	Why is the baby wearing a snowsuit?	Why does the boy have his arms in that position?	Why are two of the giraffes so much shorter than the other three?
<i>Ours:</i> surfing	cold	balance	they are babies
<i>Vgg+LSTM:</i> beach	safety	to catch	yes
<i>Ground Truth:</i> surfing	cold	balance	babies
			
Why do these sheep have paint on them?	Why is this ground white?	Why is there sand around the orange object?	Why is the man wearing black there?
<i>Ours:</i> identification	snow	safety	umpire
<i>Vgg+LSTM:</i> to eat	cold	to balance	safety
<i>Ground Truth:</i> identification	snow	safety	umpire
			
Why is she wearing a potholder on her arm?	Why is the road closed?	Why are there no leaves on the trees?	Why does he have glasses on?
<i>Ours:</i> cooking	train	winter	to see
<i>Vgg+LSTM:</i> drinking	stop	unknown	to be
<i>Ground Truth:</i> cooking	train	winter	to see
			
Why is the ground wet?	Why are the animals laying here?	Why is the man running?	Why is the cat sitting on the bench?
<i>Ours:</i> rain	resting	playing frisbee	resting
<i>Vgg+LSTM:</i> cold	no	running	to sleep
<i>Ground Truth:</i> rain	resting	playing frisbee	resting
			
Why are her hands in the air?	Why is the man standing?	Why is the child running?	Why is there a giraffe in this setting?
<i>Ours:</i> flying kite	playing tennis	flying kite	zoo
<i>Vgg+LSTM:</i> surfing	tennis ball	playing frisbee	to eat
<i>Ground Truth:</i> flying kite	playing tennis	flying kite	zoo

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.

			
What kind of weather it is?	What is in the cup?	What kind of room is this?	Where did the water come from?
<i>Ours:</i> sunny	coffee	bedroom	ocean
<i>Vgg+LSTM:</i> cloudy	wine	living	beach
<i>Ground Truth:</i> sunny	coffee	bedroom	ocean

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

			
What are the children holding?	Where is this picture?	What kind of cheese is this?	What room is this?
<i>Ours:</i> teddy bears	market	mozzarella	bathroom
<i>Vgg+LSTM:</i> wii	on left	chicken	kitchen
<i>Ground Truth:</i> teddy bears	market	mozzarella	bathroom

			
What style of cooking is this?	Is this a men's room or a women's room?	What is on the top of the animals' heads?	Is this a vegetable?
<i>Ours:</i> chinese	men's	horns	yes
<i>Vgg+LSTM:</i> pizza	hotel	rocks	no
<i>Ground Truth:</i> chinese	men's	horns	yes





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

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.





































			
What is he looking at?	What kind of meat is on this?	Which game is being played?	What brand is the bat bag?
<i>Ours:</i> toothbrush	<i>Ours:</i> bacon	<i>Ours:</i> soccer	<i>Ours:</i> nike
<i>Vgg+LSTM:</i> camera	<i>Vgg+LSTM:</i> chicken	<i>Vgg+LSTM:</i> tennis	<i>Vgg+LSTM:</i> wilson
<i>Ground Truth:</i> toothbrush	<i>Ground Truth:</i> bacon	<i>Ground Truth:</i> soccer	<i>Ground Truth:</i> nike
			
Is this inside?	What season does it look like?	Is this a healthy breakfast?	Is this meal healthy?
<i>Ours:</i> no	<i>Ours:</i> fall	<i>Ours:</i> no	<i>Ours:</i> yes
<i>Vgg+LSTM:</i> yes	<i>Vgg+LSTM:</i> winter	<i>Vgg+LSTM:</i> yes	<i>Vgg+LSTM:</i> no
<i>Ground Truth:</i> no	<i>Ground Truth:</i> fall	<i>Ground Truth:</i> no	<i>Ground Truth:</i> yes
			
The green item on the pizza, what is it called?	Does this animal have fur?	Is this a home office?	What letter is inside the blue circle?
<i>Ours:</i> broccoli	<i>Ours:</i> no	<i>Ours:</i> yes	<i>Ours:</i> p
<i>Vgg+LSTM:</i> carrots	<i>Vgg+LSTM:</i> yes	<i>Vgg+LSTM:</i> no	<i>Vgg+LSTM:</i> b
<i>Ground Truth:</i> broccoli	<i>Ground Truth:</i> no	<i>Ground Truth:</i> yes	<i>Ground Truth:</i> p
			
What type of food is this person eating?	In what type of establishment is this taken?	Is that meat on the plate?	What is being celebrated?
<i>Ours:</i> donut	<i>Ours:</i> zoo	<i>Ours:</i> yes	<i>Ours:</i> birthday
<i>Vgg+LSTM:</i> pizza	<i>Vgg+LSTM:</i> zebra	<i>Vgg+LSTM:</i> no	<i>Vgg+LSTM:</i> pizza
<i>Ground Truth:</i> donut	<i>Ground Truth:</i> zoo	<i>Ground Truth:</i> yes	<i>Ground Truth:</i> birthday
			
What kind of building is this?	Is the weather cold or warm?	Is that normal a banana on a record?	What color is the snow?
<i>Ours:</i> barn	<i>Ours:</i> cold	<i>Ours:</i> no	<i>Ours:</i> white
<i>Vgg+LSTM:</i> church	<i>Vgg+LSTM:</i> sunny	<i>Vgg+LSTM:</i> yes	<i>Vgg+LSTM:</i> blue
<i>Ground Truth:</i> barn	<i>Ground Truth:</i> cold	<i>Ground Truth:</i> no	<i>Ground Truth:</i> white

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.

			
What season is it?	What shoe company is advertised?	Is this guy going to jump high?	What is on the keyboard?
<i>Ours:</i> fall	<i>Ours:</i> vans	<i>Ours:</i> no	<i>Ours:</i> mouse
<i>Ground Truth:</i> winter	<i>Ground Truth:</i> nike	<i>Ground Truth:</i> yes	<i>Ground Truth:</i> cat

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

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

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

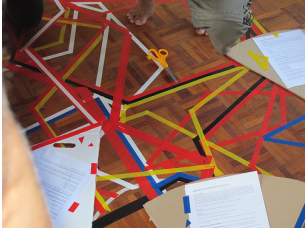



			
What's on the floor?	Who took this photo?	What food is this, really?	What kind of weather is this?
<i>Ours:</i> scissors	<i>Ours:</i> photographer	<i>Ours:</i> chicken	<i>Ours:</i> rainy
<i>Ground Truth:</i> tape	<i>Ground Truth:</i> christopher brown	<i>Ground Truth:</i> cake	<i>Ground Truth:</i> cloudy

Table 6. Some failure cases for our final model