

Questions are informative

What breed of dog is this?



This question suggests:

- the animal in the scene is a *dog*
- *breed* is a property of *dog*
- all *dogs* in the scene are the same *breed*
- knowing the *breed* may be important

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

Input	Model (trained on COCO)	Output	Result (SPICE score [1])
3 questions	Seq2seq [2]	Image caption	0.140
1 question	Used directly	Image caption	0.058
Image	Neural Talk [3]	Image caption	0.194

Questions → object classification

Question	Inferred objects
What color is the bus?	
How many umbrellas are in the image?	
Is the bird sitting on a plant?	

Infer objects from questions (on COCO): **29.3%** recall, **82.4%** precision

Method for improving VQA

Training data



Q: What is under the plane?

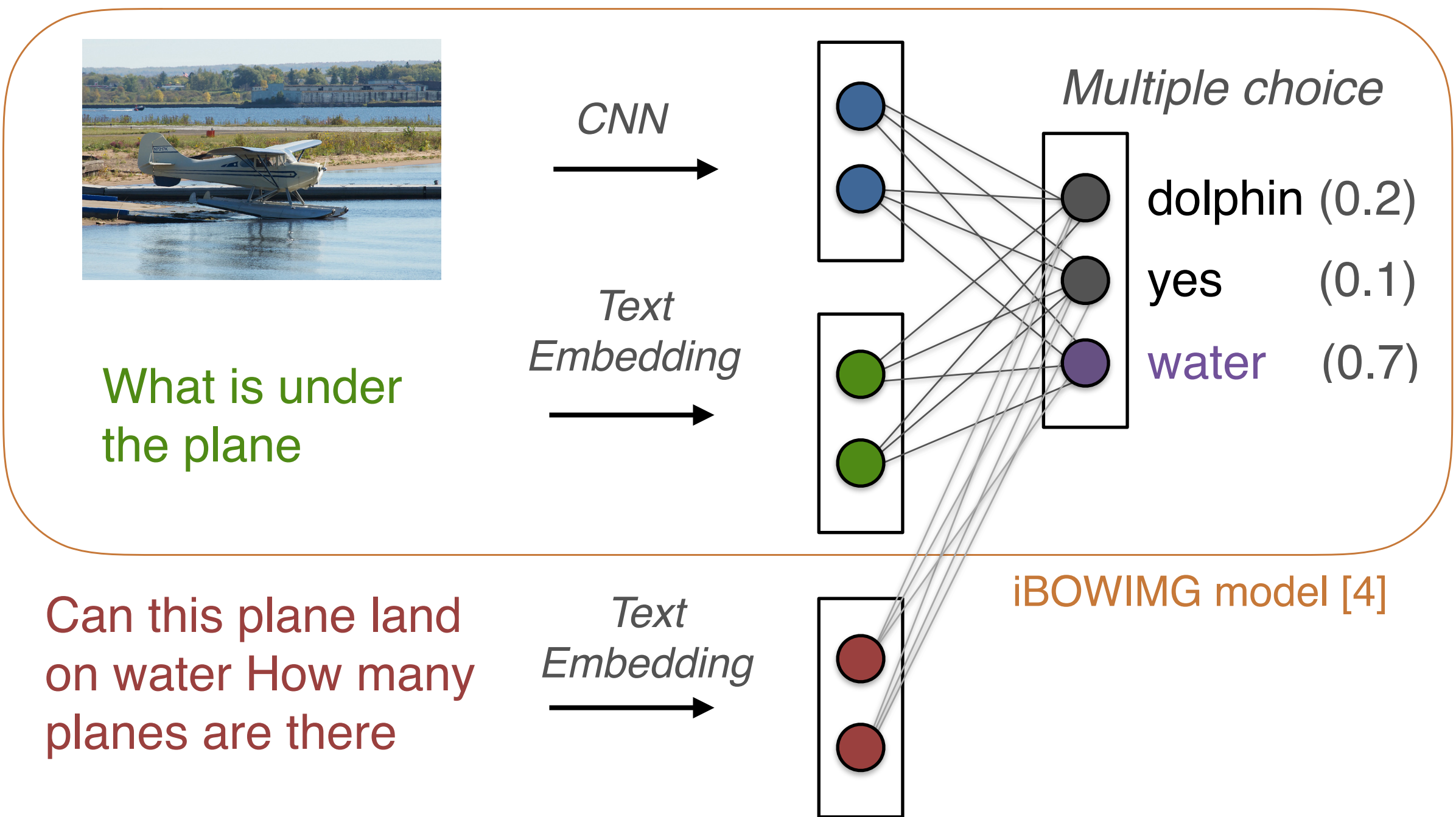
A: Water

Other questions:

Can this plane land on water?




How many planes are there?


Model: iBOWIMG-2x



Using other questions in iBOWIMG-2x

Training: ( , What is under the plane, Can this plane... are there)

Training augmented: ( , 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, \emptyset)

Testing without: ( , Is this cat lying on a sofa, \emptyset)

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

Training: Select 1 answered and leave 2 unanswered questions per training image

Test: Validation set with 3 questions per image

Model	Using unanswered questions...		Accuracy
	...at training?	...at test?	
iBOWIMG [4]	—	—	47.3
iBOWIMG-2x	yes	—	49.2
iBOWIMG-2x	yes, augmented	—	50.4
iBOWIMG-2x	yes, augmented	yes	50.9

Experiment #2: standard benchmark

Training: Train+val set with all 3 questions answered

Test: Test-dev set with 3 questions per image

Model	Accuracy	Accuracy by test question type		
		Yes/no	Number	Word
iBOWIMG [4]	55.7	76.5	34.9	42.6
iBOWIMG-2x	62.8	80.7	37.9	53.1



What object is in focus?

3
Fire Hydrant



What is in the water?

Fish
Plastic bag

Bibliography

- [1] Anderson, Fernando, Johnson and Gould. SPICE: Semantic propositional image caption evaluation. ECCV 2016
[2] Cho, Van Merrienboer, Gulcehre, Bahdanau, Bougares, Schwenk and Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. EMNLP 2014.
[3] Karpathy and Fei-Fei. Deep visual-semantic alignments for generating image descriptions. CVPR 2015
[4] Zhou, Tian, Sukhbaatar, Szlam and Fergus. Simple baseline for visual question answering. Arxiv 1512.02167, 2015.