Exploring Models and Data for Image Question Answering

Mengye Ren, Ryan Kiros, Richard S. Zemel   University of Toronto

Problem

- Image Question Answering: given an image and a free-form question, find an answer.
- We assume that answers are one-word, thus we can treat it as a classification problem.

DAQUAR 1553
What is there in front of the sofa?
Ground truth: table
I+BOW: table (0.74)
2V+BLSTM: table (0.88)
LSTM: chair (0.47)

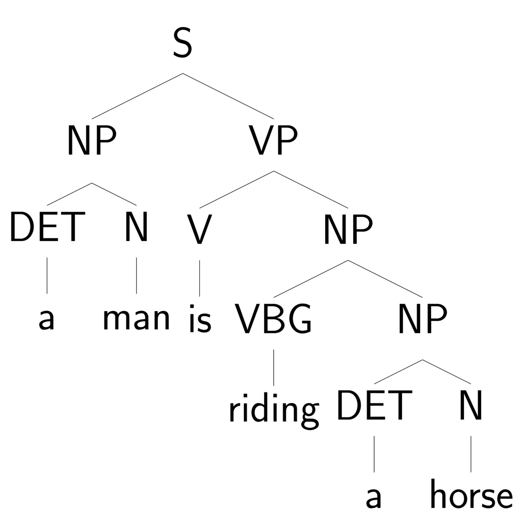
COCOQA 5078
How many leftover donuts is the red bicycle holding?
Ground truth: three
I+BOW: two (0.51)
2V+BLSTM: three (0.27)
BOW: one (0.29)

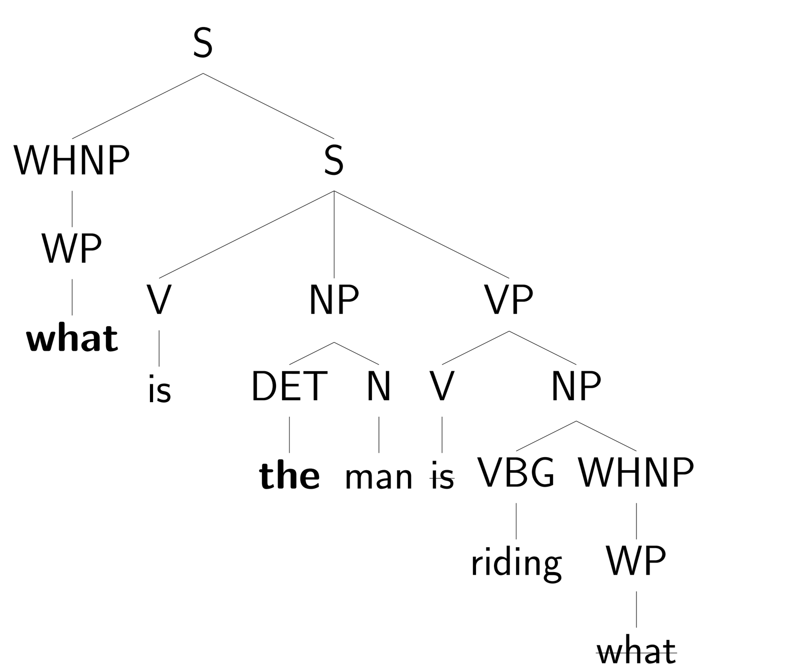
COCOQA 1238
What is the color of the tee-shirt?
Ground truth: blue
I+BOW: blue (0.31)
2V+BLSTM: orange (0.43)
BOW: green (0.38)

COCOQA 26088
Where is the gray cat sitting?
Ground truth: window
I+BOW: window (0.78)
2V+BLSTM: window (0.68)
BOW: suitcase (0.31)

Automatic QA Generation

- We generate 4 types of questions: object, number, color, location, directly from image description.
- All answers are one-word.
- We move the wh-word and verb to the front under certain constraints. For example, "A man is riding a **horse**" => "**What** is **the** man riding?"





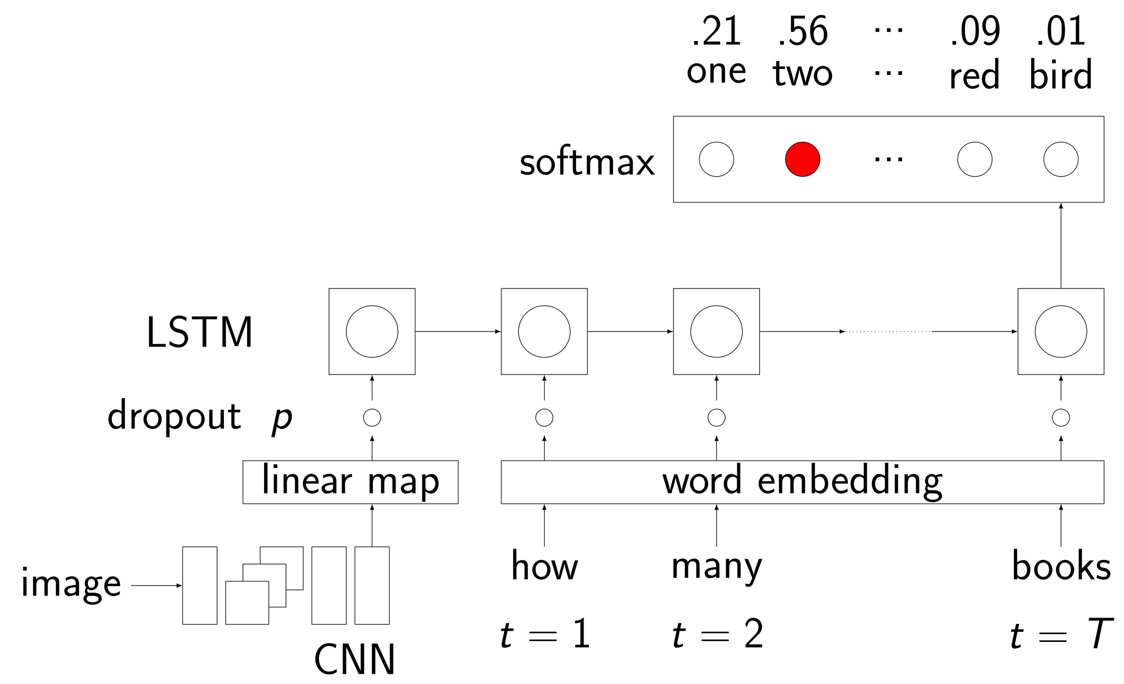
- We pruned answers that appear too rarely or too often.

Baselines

- **GUESS** Predict the most frequent answer based on the question type.
- **BOW** Given only the questions without the images and perform logistic regression on the bag-of-words vector to classify answers.
- **LSTM** Input the question words into the LSTM alone.
- **IMG** Re-train a separate CNN classification layer for each type of question.
- **IMG+PRIOR** Combine the prior knowledge of an object and the image understanding from the "deaf model". c: color, o: object of interest, and x: image. Assuming o and x are conditionally independent given the c. Use the output of the IMG model as $p(c|x)$. $p(c|o,x) = \frac{p(o|c)p(c|x)}{\sum_{c \in \mathcal{C}} p(o|c)p(c|x)}$. Empirical estimate: $\hat{p}(o|c) = \frac{count(o,c)}{count(c)}$ and Laplace smoothing.

Our Models

- **VIS+LSTM** Treats the image as a word (Vinyals et al. [1]).



- **2-VIS+BLSTM** Two image feature inputs, at the start and the end of the question, with different linear transformations. LSTMs going in forward and backward directions.
- **IMG+BOW** Softmax on top CNN features and learned BOW vectors.
- **FULL** A simple average of the three models above.

Experimental Results

Table : COCO-QA question type break-down

Category	Train	%	Test	%
Object	54992	69.84%	27206	69.85%
Number	5885	7.47%	2755	7.07%
Color	13059	16.59%	6509	16.71%
Location	4800	6.10%	2478	6.36%
Total	78736	100.00%	38948	100.00%

Table : DAQUAR and COCO-QA results

	DAQUAR			COCO-QA		
	Acc.	WUPS 0.9	WUPS 0.0	Acc.	WUPS 0.9	WUPS 0.0
MULTI-WORLD [2]	0.1273	0.1810	0.5147	-	-	-
GUESS	0.1824	0.2965	0.7759	0.0665	0.1742	0.7344
BOW	0.3267	0.4319	0.8130	0.3752	0.4854	0.8278
LSTM	0.3273	0.4350	0.8162	0.3676	0.4758	0.8234
IMG	-	-	-	0.4302	0.5864	0.8585
IMG+PRIOR	-	-	-	0.4466	0.6020	0.8624
K-NN(K=31,13)	0.3185	0.4242	0.8063	0.4496	0.5698	0.8557
IMG+BOW	0.3417	0.4499	0.8148	**0.5592**	**0.6678**	**0.8899**
VIS+LSTM	0.3441	0.4605	**0.8223**	0.5331	0.6391	0.8825
ASK-NEURON [3]	0.3468	0.4076	0.7954	-	-	-
2-VIS+BLSTM	**0.3578**	**0.4683**	0.8215	0.5509	0.6534	0.8864
FULL	**0.3694**	**0.4815**	**0.8268**	**0.5784**	**0.6790**	**0.8952**
HUMAN	0.6027	0.6104	0.7896	-	-	-

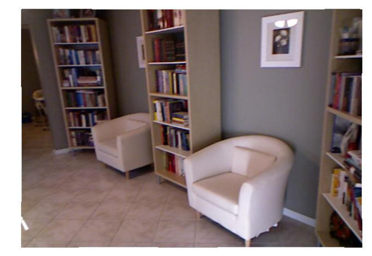
Table : COCO-QA accuracy per category

	Object	Number	Color	Location
GUESS	0.0211	0.3584	0.1387	0.0893
BOW	0.3727	0.4356	0.3475	0.4084
LSTM	0.3587	0.4534	0.3626	0.3842
IMG	0.4073	0.2926	0.4268	0.4419
IMG+PRIOR	-	0.3739	0.4899	0.4451
K-NN	0.4799	0.3699	0.3723	0.4080
IMG+BOW	**0.5866**	0.4410	**0.5196**	**0.4939**
VIS+LSTM	0.5653	**0.4610**	0.4587	0.4552
2-VIS+BLSTM	0.5817	0.4479	0.4953	0.4734
FULL	**0.6108**	**0.4766**	0.5148	**0.5028**

More Examples

COCOQA 33827
What is the color of the cat?
Ground truth: black
I+BOW: black (0.55)
2V+LSTM: black (0.73)
BOW: gray (0.40)

COCOQA 33827a
What is the color of the couch?
Ground truth: red
I+BOW: red (0.65)
2V+LSTM: black (0.44)
BOW: red (0.39)



DAQUAR 1522
How many chairs are there?
Ground truth: two
I+BOW: four (0.24)
2V+BLSTM: one (0.29)
LSTM: four (0.19)

DAQUAR 1520
How many shelves are there?
Ground truth: three
I+BOW: three (0.25)
2V+BLSTM: two (0.48)
LSTM: two (0.21)

COCOQA 14855
Where are the ripe bananas sitting?
Ground truth: basket
I+BOW: basket (0.97)
2V+BLSTM: basket (0.58)
BOW: bowl (0.48)

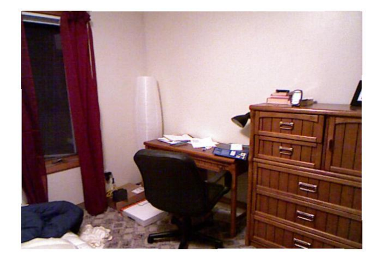
COCOQA 14855a
What are in the basket?
Ground truth: bananas
I+BOW: bananas (0.98)
2V+BLSTM: bananas (0.68)
BOW: bananas (0.14)

DAQUAR 585
What is the object on the chair?
Ground truth: pillow
I+BOW: clothes (0.37)
2V+BLSTM: pillow (0.65)
LSTM: clothes (0.40)

DAQUAR 585a
Where is the pillow found?
Ground truth: chair
I+BOW: bed (0.13)
2V+BLSTM: chair (0.17)
LSTM: cabinet (0.79)

COCOQA 23419
What is the black and white cat wearing?
Ground truth: hat
I+BOW: hat (0.50)
2V+BLSTM: tie (0.34)
BOW: tie (0.60)

COCOQA 23419a
What is wearing a hat?
Ground truth: cat
I+BOW: cat (0.94)
2V+BLSTM: cat (0.90)
BOW: dog (0.42)



DAQUAR 2136
What is right of table?
Ground truth: shelves
I+BOW: shelves (0.33)
2V+BLSTM: shelves (0.28)
LSTM: shelves (0.20)

DAQUAR 2136a
What is in front of table?
Ground truth: chair
I+BOW: chair (0.64)
2V+BLSTM: chair (0.31)
LSTM: chair (0.37)

COCOQA 22891
What is the color of the coat?
Ground truth: yellow
I+BOW: black(0.45)
V+LSTM: yellow (0.24)
BOW: red (0.28)

COCOQA 22891a
What is the color of the umbrella?
Ground truth: red
I+BOW: black (0.28)
V+LSTM: yellow (0.26)
BOW: red (0.29)

COCOQA 498
How many vintage refrigerators blue and red in color?
Ground truth: four
I+BOW: five (0.25)
2V+BLSTM: four (0.29)
BOW: one (0.24)

COCOQA 498a
How many refrigerators are blue?
Ground truth: two
I+BOW: four (0.35)
2V+BLSTM: six (0.09)
BOW: two (0.37)

Figure : Sample questions and responses of our system. For some of the examples, we specifically tested extra questions (the ones have an "a" in the question ID).

Download

- Download dataset, software (models and question generation), and full results at **http://www.cs.toronto.edu/~mren/imageqa**

Conclusion

- We present our end-to-end neural network models to image QA.
- Simple bag-of-words can perform equally well compared to recurrent neural network.
- Models have large space for improvement on questions such as color and counting.
- We release an Image QA dataset that is automatically generated from image description.

References

[1] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In CVPR, 2015.

[2] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In NIPS, 2014.

[3] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask Your Neurons: A Neural-based Approach to Answering Questions about Images. In ICCV, 2015.