Exploring Models and Data for Image Question Answering

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

Computer Science UNIVERSITY OF TORONTO

Problem

- Image Question Answering (QA): 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
IMG+BOW: table (0.74)
2-VIS+BLSTM: table (0.88)
LSTM: chair (0.47)

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

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

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

Figure: Sample questions and responses of a variety of models. Correct answers are in green and incorrect in red. The numbers in parentheses are the probabilities assigned to the top-ranked answer by the given model. The leftmost example is from the DAQUAR dataset, and the others are from our new COCO-QA datset.

Our Models

- **VIS+LSTM Model**

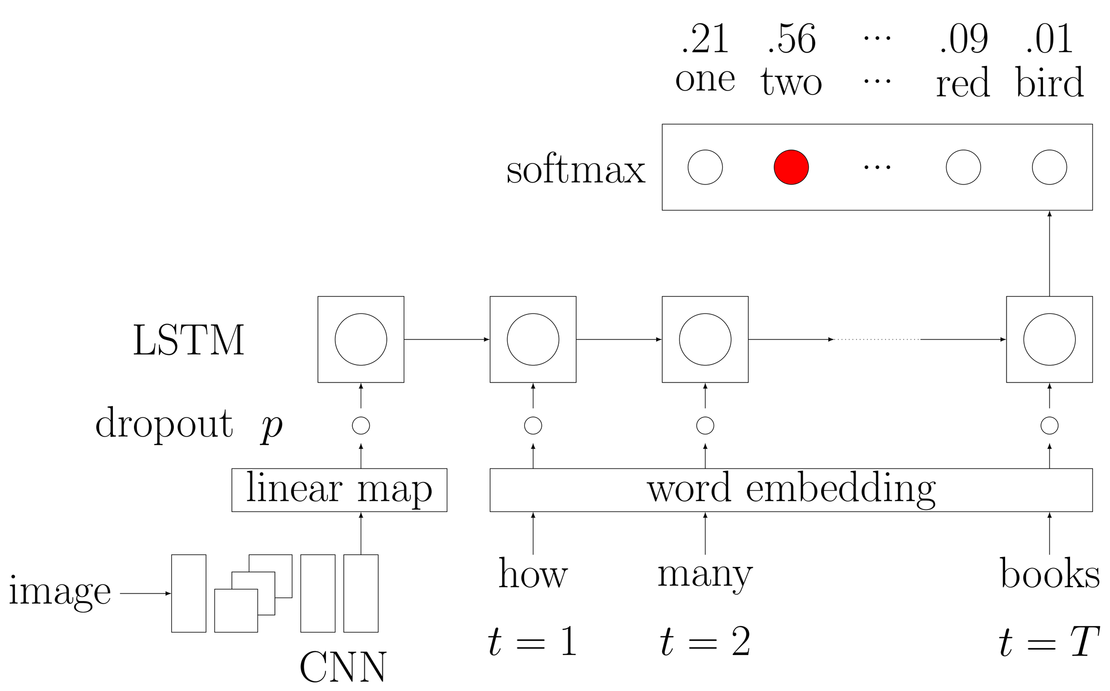


Figure: VIS+LSTM Model

  - Borrowed the idea of treating the image as a word from previous caption generation work done by Vinyals et al. [1].
  - Image features are passed through a linear transformation to match with word vector dimensions.
  - At the last time step, the Long Short Term Memory (LSTM) [2] outputs to the softmax layer to classify answers.
- **2-VIS+BLSTM Model**
  - Two image feature inputs, at the start and the end of the question, with different learned linear transformations.
  - LSTMs going in both the forward and backward directions.
  - Both LSTMs output to the softmax layer.
- **IMG+BOW** Multinomial logistic regression based on the CNN image features (4096 dimension), and learned bag-of-word (BOW) vectors.
- **FULL** A simple average of the three models above.

Automatic QA Generation

- **Motivation** Currently available dataset DAQUAR [3] is very small (1500 images, 7000 QA on 37 classes of objects, 12000 QA on 894 classes of objects). Guessing the modes can yield very good accuracy, and "blind models" can achieve almost equal performance compared the best model.
- 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 mode 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". Denote c as the color, o as the class of the object of interest, and x as the image. Assuming o and x are conditionally independent given the color. 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.

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

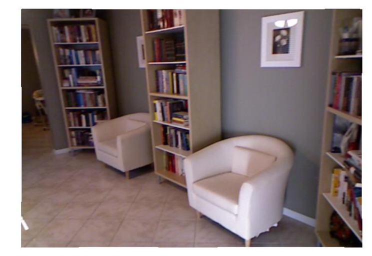
	DAQUAR			COCO-QA		
	Acc.	WUPS 0.9	WUPS 0.0	Acc.	WUPS 0.9	WUPS 0.0
MULTI-WORLD [4]	0.1273	0.1810	0.5147	-	-	-
GUESS	0.1824	0.2965	0.7759	0.0665	0.1742	0.7344
BOW	0.3267	0.4319	0.8130	0.3752	0.4854	0.8278
LSTM	0.3273	0.4350	0.8162	0.3676	0.4758	0.8234
IMG	-	-	-	0.4302	0.5864	0.8585
IMG+PRIOR	-	-	-	0.4466	0.6020	0.8624
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 [5]	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

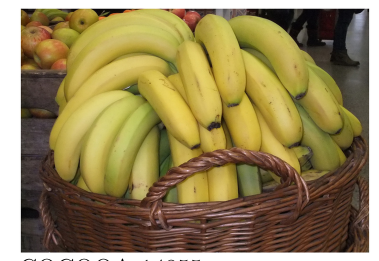
	Object	Number	Color	Location
GUESS	0.0211	0.3584	0.1387	0.0893
BOW	0.3727	0.4356	0.3475	0.4084
LSTM	0.3587	0.4534	0.3626	0.3842
IMG	0.4073	0.2926	0.4268	0.4419
IMG+PRIOR	-	0.3739	0.4899	0.4451
IMG+BOW	**0.5866**	0.4410	**0.5196**	**0.4939**
VIS+LSTM	0.5653	**0.4610**	0.4587	0.4552
2-VIS+BLSTM	0.5817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

More Examples

COCOQA 4018
What is the color of the bowl?
Ground truth: blue
IMG+BOW: blue (0.49)
2-VIS+LSTM: blue (0.52)
BOW: white (0.45)
COCOQA 4018a
What is the color of the vest?
Ground truth: red
IMG+BOW: red (0.29)
2-VIS+LSTM: orange (0.37)
BOW: orange (0.57)

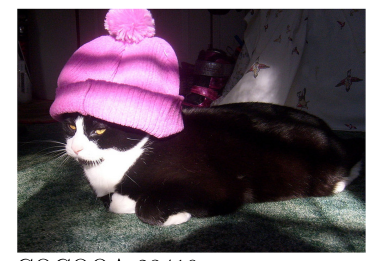


DAQUAR 1522
How many chairs are there?
Ground truth: two
IMG+BOW: four (0.24)
2-VIS+BLSTM: one (0.29)
LSTM: four (0.19)
DAQUAR 1520
How many shelves are there?
Ground truth: three
IMG+BOW: three (0.25)
2-VIS+BLSTM: two (0.48)
LSTM: two (0.21)

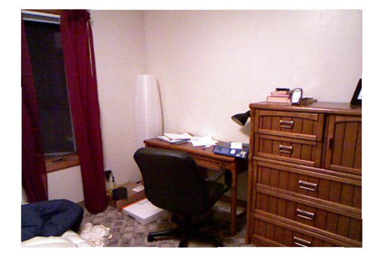


COCOQA 14855
Where are the ripe bananas sitting?
Ground truth: basket
IMG+BOW: basket (0.97)
2-VIS+BLSTM: basket (0.58)
BOW: bowl (0.48)
COCOQA 14855a
What are in the basket?
Ground truth: bananas
IMG+BOW: bananas (0.98)
2-VIS+BLSTM: bananas (0.68)
BOW: bananas (0.14)

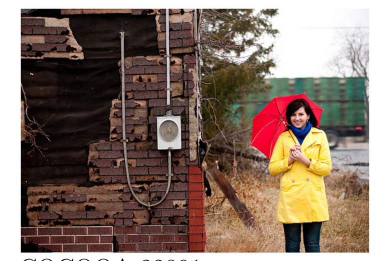
DAQUAR 585
What is the object on the chair?
Ground truth: pillow
IMG+BOW: clothes (0.37)
2-VIS+BLSTM: pillow (0.65)
LSTM: clothes (0.40)
DAQUAR 585a
Where is the pillow found?
Ground truth: chair
IMG+BOW: bed (0.13)
2-VIS+BLSTM: chair (0.17)
LSTM: cabinet (0.79)



COCOQA 23419
What is the black and white cat wearing?
Ground truth: hat
IMG+BOW: hat (0.50)
2-VIS+BLSTM: tie (0.34)
BOW: tie (0.60)
COCOQA 23419a
What is wearing a hat?
Ground truth: cat
IMG+BOW: cat (0.94)
2-VIS+BLSTM: cat (0.90)
BOW: dog (0.42)



DAQUAR 2136
What is right of table?
Ground truth: shelves
IMG+BOW: shelves (0.33)
2-VIS+BLSTM: shelves (0.28)
LSTM: shelves (0.20)
DAQUAR 2136a
What is in front of table?
Ground truth: chair
IMG+BOW: chair (0.64)
2-VIS+BLSTM: chair (0.31)
LSTM: chair (0.37)



COCOQA 22891
What is the color of the coat?
Ground truth: yellow
IMG+BOW: black(0.45)
VIS+LSTM: yellow (0.24)
BOW: red (0.28)
COCOQA 22891a
What is the color of the umbrella?
Ground truth: red
IMG+BOW: black (0.28)
VIS+LSTM: yellow (0.26)
BOW: red (0.29)

COCOQA 498
How many vintage refrigerators blue and red in color?
Ground truth: four
IMG+BOW: five (0.25)
2-VIS+BLSTM: four (0.29)
BOW: one (0.24)
COCOQA 498a
How many refrigerators are blue?
Ground truth: two
IMG+BOW: four (0.35)
2-VIS+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).

- Full results: http://www.cs.toronto.edu/~mren/imageqa/results

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. Download: http://www.cs.toronto.edu/~mren/imageqa/data/cocoqa

Current Directions

- Free-form text generation model. Similar to image captioning, it will require an automatic free-form answer evaluation metric.
- Extend questions to open domain.
- Use of visual attention to improve results and interpret output (based on recent successes of visual attention in image captioning [6]).

References

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

[2] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[3] Mateusz Malinowski and Mario Fritz. Towards a visual turing challenge. In *NIPS Workshop on Learning Semantics*, 2014.

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

[5] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask Your Neurons: A Neural-based Approach to Answering Questions about Images. *CoRR*, abs/1505.01121, 2015.

[6] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML (to appear)*, 2015.