Exploring Models and Data for Image Question Answering

Mengye Ren[1]   MREN@CS.TORONTO.EDU
Ryan Kiros[1]   RKIROS@CS.TORONTO.EDU
Richard Zemel[1,2]   ZEMEL@CS.TORONTO.EDU

[1]Department of Computer Science, University of Toronto, Toronto, ON, CANADA
[2]Canadian Institute for Advanced Research, Toronto, ON, CANADA

Abstract

This work aims to address the problem of image-based question-answering (QA) with new models and datasets. In our work, we propose to use neural networks and visual semantic embeddings, without intermediate stages such as object detection and image segmentation, to predict answers to simple questions about images. Our model performs 1.8 times better than the only published results on an existing image QA dataset. We also present a question generation algorithm that converts image descriptions, which are widely available, into QA form. We used this algorithm to produce an order-of-magnitude larger dataset, with more evenly distributed answers. A suite of baseline results on this new dataset are also presented.

1. Introduction

Combining image understanding and natural language interaction is one of the grand dreams of artificial intelligence. We are interested in the problem of jointly learning image and text through a question-answering task. Recently, researchers studying image caption generation (Vinyals et al., 2015; Kiros et al., 2015; Karpathy et al., 2013; Mao et al., 2014; Donahue et al., 2014; Chen & Zitnick, 2014; Fang et al., 2015; Xu et al., 2015; Lebret et al., 2015; Klein et al., 2015) have developed powerful methods of jointly learning from image and text inputs to form higher level representations from models such as convolutional neural networks (CNNs) trained on object recognition, and word embeddings trained on large scale text corpora. Image QA involves an extra layer of interation between human and computers. Here the model needs to pay attention to details of the image instead of describing it in

a vague sense. The problem also combines many computer vision sub-problems such as image labelling and object detection.

In this paper we present our contributions to the problem: a generic end-to-end QA model using visual semantic embeddings to connect a CNN and a recurrent neural net (RNN), as well as comparisons to a suite of other models; an automatic question generation algorithm that converts description sentences into questions; and a new QA dataset (COCO-QA) that was generated using the algorithm, and a number of baseline results on this new dataset.

2. Problem Formulation

The inputs of the problem are an image and a question, and the output is an answer. In this work we assume that the answers consist of only a single word, which allows us to treat the problem as a classification problem. This also makes the evaluation of the models easier and more robust, avoiding the thorny evaluation issues that plague multi-word generation problems.

3. Related Work

Malinowski & Fritz (2014a) released a dataset with images and question-answer pairs, the DAtaset for QUestion Answering on Real-world images (DAQUAR). All images are from the NYU depth v2 dataset (Silberman et al., 2012), and are taken from indoor scenes. Human segmentations, image depth values, and object labellings are available in the dataset. The QA data has two sets of configurations, which differ by the number of object classes appearing in the questions (37-class and 894-class). There are mainly three types of questions in this dataset: object type, object color, and number of objects. Some questions are easy but many questions are very hard to answer even for humans. Since DAQUAR is the only publicly available image-based QA dataset, it is one of our benchmarks to evaluate our models.

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 1. 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.

Together with the release of the DAQUAR dataset, Malinowski et al. presented an approach which combines semantic parsing and image segmentation. In the natural language part of their work, they used a semantic parser (Liang et al., 2013) to convert sentences into latent logical forms. They obtained multiple segmentations of the image by sampling the uncertainty of the segmentation algorithm. Lastly they used a Bayesian approach to sample from the nearest neighbors in the training set according to the similarity of the predicates.

This approach is notable as one of the first attempts at image QA, but it has a number of limitations. First, a human-defined possible set of predicates are very dataset-specific. To obtain the predicates, their algorithm also depends on the accuracy of the image segmentation algorithm and image depth information. Second, before asking any of the questions, their model needs to compute all possible spatial relations in the training images. Even though the model searches from the nearest neighbors of the test images, it could still be an expensive operation in larger datasets. Lastly the accuracy of their model is not very strong. We will show later that some simple baselines will perform bet-

ter in terms of plain accuracy.

Very recently there has been a number of parallel efforts on both creating datasets and proposing new models (Antol et al., 2015; Malinowski et al., 2015; Gao et al., 2015; Ma et al., 2015). Both Antol et al. (2015) and Gao et al. (2015) used MS-COCO (Lin et al., 2014) images and created an open domain dataset with human generated questions and answers. In Anto et al.'s work, the authors also included cartoon pictures besides real images. Some questions require logical reasoning in order to answer correctly.

Both Malinowski et al. (2015) and Gao et al. (2015) use recurrent networks to encode the sentence and output the answer. Whereas Malinowski et al. use a single network to handle both encoding and decoding, Gao et al. used two networks, a separate encoder and decoder. Lastly, bilingual (Chinese and English) versions of the QA dataset are available in Gao et al.'s work. Ma et al. (2015) used convolutional neural networks (CNNs) to both extract image features and sentence features, and fuse the features together with another multi-modal CNN.

Our approach is developed independently from the work above. Similar to the work of Malinowski et al. (2015) and Gao et al. (2015), we also experimented with recurrent networks to consume the sequential question input. Unlike Gao et al. (2015), we formulate the task as a classification problem, as there is no single well-accepted metric to evaluate sentence-form answer accuracy (Chen et al., 2015). Thus, we place more focus on a limited domain of questions that can be answered with one word. We also formuate and evaluate a range of other algorithms, that utilize various representations drawn from the question and image, on these datasets.

4. Proposed Methodology

The methodology presented here is two-fold. On the model side we develop and apply various forms of neural networks and visual-semantic embeddings on this task, and on the dataset side we propose new ways of synthesizing QA pairs from currently available image description datasets.

4.1. Models

In recent years, recurrent neural networks (RNNs) have enjoyed some successes in the field of natural language processing (NLP). Long short-term memory (LSTM) (Hochreiter & Schmidhuber, 1997) is a form of RNN which is easier to train than standard RNNs because of its linear error propagation and multiplicative gatings. There has been increasing interest in using LSTM as encoders and decoders on sentence level. Our model builds directly on top of the LSTM sentence model and is called the "VIS+LSTM" model. It treats the image as one word of the question. We
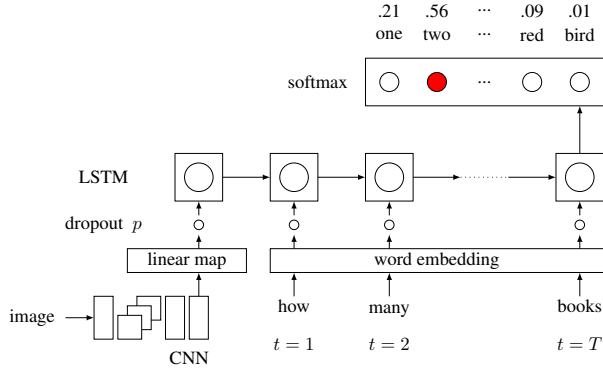
Figure 2. VIS+LSTM Model

borrowed this idea of treating the image as a word from caption generation work done by Vinyals et al. (2015). We will compare this newly proposed model with a suite of simpler models in the Experimental Results section.

1. We use the last hidden layer of the 19-layer Oxford VGG Conv Net (Simonyan & Zisserman, 2015) trained on ImageNet 2014 Challenge (Russakovsky et al., 2015) as our visual embeddings. The conv-net (CNN) part of our model is kept frozen during training.

2. We experimented with several different word embedding models: randomly initialized embedding, dataset-specific skip-gram embedding and general-purpose skip-gram embedding model (Mikolov et al., 2013). The word embeddings can either be frozen or dynamic (trained with the rest of the model).

3. We then treat the image as if it is the first word of the sentence. Similar to DeViSE (Frome et al., 2013), we use a linear or affine transformation to map 4096 dimension image feature vectors to a 300 or 500 dimensional vector that matches the dimension of the word embeddings.

4. We can optionally treat the image as the last word of the question as well through a different weight matrix.

5. We can optionally add a reverse LSTM, which gets the same content but operates in a backward sequential fashion.

6. The LSTM(s) outputs are fed into a softmax layer at the last timestep to generate answers.

4.2. Question-Answer Generation

The currently available DAQUAR dataset contains approximately 1500 images and 7000 questions on 37 common object classes, which might be not enough for training large complex models. Another problem with the current dataset is that simply guessing the modes can yield very good accuracy.

We aim to create another dataset, to produce a much larger number of QA pairs and a more even distribution of answers. While collecting human generated QA pairs is one possible approach, and another is to synthesize questions based on image labellings, we instead propose to automatically convert descriptions into QA form. In general, objects mentioned in image descriptions are easier to notice than the ones in DAQUAR's human generated questions, and than the ones in synthetic QAs based on ground truth labellings. This allows the model to rely more on rough image understanding without any logical reasoning. Lastly the conversion process preserves the language variability in the original description, and results in more human-like questions than questions generated from image labellings.

Question generation is still an open-ended topic. As a starting point we used the MS-COCO dataset (Lin et al., 2014), but the same method can be applied to any other image description dataset, such as Flickr (Hodosh et al., 2013), SBU (Ordonez et al., 2011), or even the internet. We adopt a conservative approach to generating questions in an attempt to create high-quality questions.

4.2.1. PRE-PROCESSING & COMMON STRATEGIES

We used the Stanford parser (Klein & Manning, 2003) to obtain the syntatic structure of the original image description. We also utilized these strategies for forming the questions.

1. Compound sentences to simple sentences
   Here we only consider a simple case, where two sentences are joined together with a conjunctive word. We split the orginial sentences into two independent sentences.

2. Indefinite determiners "a(n)" to definite determiners "the".

3. Wh-movement constraints
   In English, questions tend to start with interrogative words such as "what". The algorithm needs to move the verb as well as the "wh-" constituent to the front of the sentence. For example: "A man is riding a **horse**" becomes "**What** is **the** man riding?" In this work we consider the following two simple constraints: (1) A-over-A principle which restricts the movement of a wh-word inside a noun phrase (NP) (Chomsky, 1973); (2) Our algorithm does not move any wh-word that is contained in a clause constituent.

4.2.2. QUESTION GENERATION

1. *Object* **Questions**: First, we consider asking about an object using "what". This involves replacing the actual object with a "what" in the sentence, and then transforming the sentence structure so that the "what" appears in the front of the sentence. The entire algorithm has the following stages: (1) Split long sentences into simple sentences; (2) Change indefinite determiners to definite determiners; (3) Traverse the sentence and identify potential answers and replace with "what". During the traversal of object-type question generation, we currently ignore all the prepositional phrase (PP) consituents; (4) Perform wh-movement. In order to identify a possible answer word, we used WordNet (Fellbaum, 1998) and the NLTK software package (Bird, 2006) to get noun categories.

2. *Number* **Questions**: We follow a similar procedure as the previous algorithm, except for a different way to identify potential answers: we extract numbers from original sentences. Splitting compound sentences, changing determiners, and wh-movement parts remain the same.

3. *Color* **Questions**: Color questions are much easier to generate. This only requires locating the color adjective and the noun to which the adjective attaches. Then it simply forms a sentence "What is the color of the [object]" with the "object" replaced by the actual noun.

4. *Location* **Questions**: These are similar to generating object questions, except that now the answer traversal will only search within PP contituents that start with the preposition "in". We also added rules to filter out clothing so that the answers will mostly be places, scenes, or large objects that contain smaller objects.

4.2.3. POST-PROCESSING

We rejected the answers that appear too rarely or too often in our generated dataset. Details can be found in Supplementary Materials. After this QA rejection process the mode composition reduced from 24.80% down to 6.65% in the test set of COCO-QA.

5. Experimental Results

5.1. Datasets

Table 1 summarizes the statistics of COCO-QA. It should be noted that since we applied the QA pair rejection process, mode-guessing performs very poorly on COCO-QA. However, COCO-QA questions are actually easier to answer than DAQUAR from a human point of view. This encourages the model to exploit salient

Table 1. 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%

object relations instead of exhaustively searching all possible relations. COCO-QA dataset can be downloaded at `http://www.cs.toronto.edu/~mren/imageqa/data/cocoqa`

5.2. Model Details

1. **VIS+LSTM**: The first model is the CNN and LSTM with a dimensionality-reduction weight matrix in the middle; we call this "VIS+LSTM" in our tables and figures.

2. **2-VIS+BLSTM**: The second model has two image feature inputs, at the start and the end of the sentence, with different learned linear transformations, and also has LSTMs going in both the forward and backward directions. Both LSTMs output to the softmax layer at the last timestep. We call the second model "2-VIS+BLSTM".

3. **IMG+BOW**: This simple model performs multinomial logistic regression based on the image features without dimensionality reduction (4096 dimension), and a bag-of-word (BOW) vector obtained by summing all the word vectors of the question.

4. **FULL**: Lastly, the "FULL" model is a simple average of the three models above.

5.3. Baselines

To evaluate the effectiveness of our models, we designed a few baselines.

1. **GUESS**: One very simple baseline is to predict the mode based on the question type. For example, if the question contains "how many" then the model will output "two." In DAQUAR, the modes are "table", "two", and "white" and in COCO-QA, the modes are "cat", "two", "white", and "room".

2. **BOW**: We designed a set of "blind" models which are given only the questions without the images. One of the

simplest blind models performs logistic regression on the BOW vector to classify answers.

3. **LSTM**: Another "blind" model we experimented with simply inputs the question words into the LSTM alone.

4. **IMG**: We also trained a counterpart "deaf" model. For each type of question, we train a separate CNN classification layer (with all lower layers frozen during training). Note that this model knows the type of question, in order to make its performance somewhat comparable to the models that can take into account the words to narrow down the answer space. However the model does not know anything about the question except the type.

5. **IMG+PRIOR**: This baseline combines the prior knowledge of an object and the image understanding from the "deaf model". For example, a question asking the color of a white bird flying in the blue sky may output white rather than blue simply because the prior probability of the bird being blue is lower. We 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,

$$
\begin{aligned}
p(c|o, x) &= \frac{p(c, o|x)}{\sum_{c \in \mathcal{C}} p(c, o|x)} = \frac{p(o|c, x)p(c|x)}{\sum_{c \in \mathcal{C}} p(o|c, x)p(c|x)} \\
&= \frac{p(o|c)p(c|x)}{\sum_{c \in \mathcal{C}} p(o|c)p(c|x)}
\end{aligned}
\tag{1}
$$

This can be computed if $p(c|x)$ is the output of a logistic regression given the CNN features alone, and we simply estimate $p(o|c)$ empirically: $\hat{p}(o|c) = \frac{count(o,c)}{count(c)}$. We use Laplace smoothing on this empirical distribution.

5.4. Performance Metrics

To evaluate model performance, we used the plain answer accuracy as well as the Wu-Palmer similarity (WUPS) measure (Wu & Palmer, 1994; Malinowski & Fritz, 2014b). The WUPS calculates the similarity between two words based on their longest common subsequence in the taxonomy tree. The similarity function takes in a threshold parameter. If the similarity between two words is less than the threshold then a score of zero will be given to the candidate answer. It reduces to plain accuracy when the threshold equals to 1.0. Following Malinowski & Fritz (2014b), we measure all the models in terms of accuracy, WUPS 0.9, and WUPS 0.0.

Table 2 summarizes the learning results on DAQUAR. Here we compare our results with (Malinowski & Fritz, 2014b). It should be noted that their result is for the entire dataset whereas our result is for the 98.3% of the original dataset with single-word answers.

5.5. Performance Metrics

To evaluate model performance, we used the plain answer accuracy as well as the Wu-Palmer similarity (WUPS) measure (Wu & Palmer, 1994; Malinowski & Fritz, 2014b). The WUPS calculates the similarity between two words based on their longest common subsequence in the taxonomy tree. If the similarity between two words is less than a threshold then a score of zero will be given to the candidate answer. Following Malinowski & Fritz (2014b), we measure all the models in terms of accuracy, WUPS 0.9, and WUPS 0.0.

5.6. Results and Analysis

Table 2 summarizes the learning results on DAQUAR. Here we compare our results with (Malinowski & Fritz, 2014b) and (Malinowski et al., 2015). It should be noted that our result is for the 98.3% of the reduced dataset with single-word answers.

From the above results we observe that our model outperforms the baselines and the existing approach in terms of answer accuracy and WUPS. Our VIS+LSTM and Malinowski et al. (2015)'s recurrent neural network model achieved somewhat similar performance on DAQUAR. It is surprising to see that IMG+BOW model is very strong on both dataset. One limitation of our VIS+LSTM model is that we are not able to consume image features as large as 4096 dimensions at one time step, so the dimensionality reduction may lose some useful information. This also shows that in Image QA tasks, and in particular on the simple questions studied here, sequential word interaction may not be as important as in other natural language tasks. A simple average of all three models further boost the performance by 1-2%, outperforming other models.

It is also interesting that the blind model does not lose much on the DAQUAR dataset, We speculate that it is likely that the ImageNet images are very different from the indoor scene images which are mostly composed of furniture. However, the non-blind models outperform the blind models by a large margin on the COCO-QA. There are three possible reasons: (1) the objects in MS-COCO resemble the ones in ImageNet more; (2) MS-COCO images have fewer objects whereas the indoor scenes have considerable clutter; and (3) COCO-QA has more data to train complex models.

There are many interesting examples but due to space limitations we can only show a few in Figure 3 and more in Supplementary Materials. Full results are available to view at http://www.cs.toronto.edu/~mren/imageqa/results.

For some of the examples, we specifically tested extra questions (the ones have an "a" in the question ID); these pro-

Table 2. DAQUAR and COCO-QA results

	DAQUAR			COCO-QA		
	ACC.	WUPS 0.9	WUPS 0.0	ACC.	WUPS 0.9	WUPS 0.0
MULTI-WORLD (MALINOWSKI & FRITZ, 2014B)	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 (MALINOWSKI ET AL., 2015)	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 3. 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**

vide more insight into the models' representation of the image and question information, and help elucidate questions that our models accidentally got correct. The parentheses in the figures represent the confidence score given by the softmax layer of the models.

Model Selection: We did not find that using different word embedding has a significant impact on the final classification results. We observed that fine-tuning the word embedding results in better performance and normalizing the CNN hidden image features into zero-mean and uni-variance help achieve faster training time. The bidirectional LSTM model can further boost the result by a little.

Object Questions: As the original CNN was trained for the ImageNet challenge, the IMG+BOW benefited largely from its single object recognition ability. Usually, the IMG+BOW and VIS+LSTM can easily get the correct answer just from the image features. However, the challenging part is to consider spatial relations between multiple objects and to focus on details of the image. Some qualitative results in Figure 3 show that the models only did a moderately acceptable job on it. Sometimes it fails to

make a correct decision but outputs the most salient object, while sometimes the blind model can equally guess the most probable objects based on the question alone (e.g., chairs should be around the dining table). Nonetheless, the FULL model improves accuracy by 50% compared to IMG model, which shows the difference between pure object classification and image question answering.

Counting: In DAQUAR, we could not observe any advantage in the counting ability of the VIS+LSTM model compared to other blind baselines. In COCO-QA there is some observable counting ability in very clean images with single object type. The model can sometimes count up to five or six. However, as shown in Figure 3, the ability is fairly weak as it does not count correctly when different object types are present. There is a lot of room for improvement in the counting task, and in fact this could be a separate computer vision problem on its own.

Color: In COCO-QA there is a significant win for the IMG+BOW and the VIS+LSTM model against the blind models on color-type questions. We further discovered that the model is not only able to recognize the dominant

color of the image but sometimes associate different colors to different objects, as shown in Figure 3. However, the model still fails on a number of easy examples. Adding prior knowledge provides an immediate gain on the "IMG" model in terms of accuracy on Color and Number questions. The gap between the IMG+PRIOR and IMG+BOW shows some localized color association ability in the CNN image representation.

6. Conclusion and Current Directions

In this paper, we consider the image QA problem and present our end-to-end neural network models. Our model shows a reasonable understanding of the question and some coarse image understanding, but it is still very naive in many situations. While recurrent networks are becoming a popular choice for learning image and text, we show that a simple bag-of-words can perform equally well compared to a recurrent network that is borrowed from an image caption generation framework (Vinyals et al., 2015). We proposed a more complete set of baselines which can provide potential insight for developing more sophisticated end-to-end image question answering systems. As the currently available dataset is not large enough, we designed an algorithm that helps us collect large scale image QA dataset from image descriptions. Our question generation algorithm is extensible to many image description datasets and can be automated without requiring extensive human effort. We hope that the release of the new dataset will encourage more data-driven approaches to this problem in the future.

Image question answering is a fairly new research topic, and the approach we present here has a number of limitations. First the model is just an answer classifier. Ideally we would like to permit longer answers which will involve some sophisticated text generation model or structured output. But this will require an automatic free-form answer evaluation metric. Second, we are only focusing on a limited domain of questions. However, this limited range of questions allow us to study the results more in depth. Lastly, it is also hard to interpret why the model outputs a certain answer. By comparing the model to some baselines we can roughly infer whether the model understood the image. Visual attention is another future direction, which could both improve the results (based on recent successes in image captioning (Xu et al., 2015)) as well as help explain the model output by examining the attention output at every timestep.

Acknowledgments

References

Antol, Stanislaw, Agrawal, Aishwarya, Lu, Jiasen, Mitchell, Margaret, Batra, Dhruv, Zitnick, C. Lawrence, and Parikh, Devi. VQA: Visual Question Answering. *CoRR*, abs/1505.00468, 2015.

Bird, Steven. NLTK: the natural language toolkit. In *ACL*, 2006.

Chen, Xinlei and Zitnick, C. Lawrence. Learning a recurrent visual representation for image caption generation. *CoRR*, abs/1411.5654, 2014.

Chen, Xinlei, Fang, Hao, Lin, Tsung-Yi, Vedantam, Ramakrishna, Gupta, Saurabh, Dollár, Piotr, and Zitnick, C. Lawrence. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325, 2015.

Chomsky, Noam. *Conditions on Transformations*. Academic Press, New York, 1973.

Donahue, Jeff, Hendricks, Lisa Anne, Guadarrama, Sergio, Rohrbach, Marcus, Venugopalan, Subhashini, Saenko, Kate, and Darrell, Trevor. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2014.

Fang, Hao, Gupta, Saurabh, Iandola, Forrest N., Srivastava, Rupesh, Deng, Li, Dollár, Piotr, Gao, Jianfeng, He, Xiaodong, Mitchell, Margaret, Platt, John C., Zitnick, C. Lawrence, and Zweig, Geoffrey. From captions to visual concepts and back. In *CVPR*, 2015.

Fellbaum, Christiane (ed.). *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London, May 1998. ISBN 978-0-262-06197-1.

Frome, Andrea, Corrado, Gregory S., Shlens, Jonathon, Bengio, Samy, Dean, Jeffrey, Ranzato, Marc'Aurelio, and Mikolov, Tomas. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013.

Gao, Haoyuan, Mao, Junhua, Zhou, Jie, Huang, Zhiheng, Wang, Lei, and Xu, Wei. Are you talking to a machine? dataset and methods for multilingual image question answering. *CoRR*, abs/1505.05612, 2015.

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

Hodosh, Micah, Young, Peter, and Hockenmaier, Julia. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Intell. Res. (JAIR)*, 47:853–899, 2013.

Karpathy, Andrej, Joulin, Armand, and Fei-Fei, Li. Deep fragment embeddings for bidirectional image sentence mapping. In *NIPS*, 2013.

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)

Figure 3. Sample questions and responses of our system

Kiros, Ryan, Salakhutdinov, Ruslan, and Zemel, Richard S. Unifying visual-semantic embeddings with multimodal neural language models. *TACL*, 2015.

Klein, Benjamin, Lev, Guy, Lev, Gil, and Wolf, Lior. Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotations. In *CVPR*, 2015.

Klein, Dan and Manning, Christopher D. Accurate unlexicalized parsing. In *ACL*, 2003.

Lebret, Rémi, Pinheiro, Pedro O., and Collobert, Ronan. Phrase-based image captioning. In *ICML (to appear)*, 2015.

Liang, Percy, Jordan, Michael I., and Klein, Dan. Learning dependency-based compositional semantics. *Computational Linguistics*, 39(2):389–446, 2013.

Lin, Tsung-Yi, Maire, Michael, Belongie, Serge, Hays, James, Perona, Pietro, Ramanan, Deva, Dollár, Piotr, and Zitnick, C. Lawrence. Microsoft COCO: common objects in context. In *ECCV*, 2014.

Ma, Lin, Lu, Zhengdong, and Li, Hang. Learning to answer questions from image using convolutional neural network. *CoRR*, abs/1506.00333, 2015.

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

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

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

Mao, Junhua, Xu, Wei, Yang, Yi, Wang, Jiang, and Yuille, Alan L. Explain images with multimodal recurrent neural networks. *NIPS Deep Learning Workshop*, 2014.

Mikolov, Tomas, Chen, Kai, Corrado, Greg, and Dean, Jeffrey. Efficient estimation of word representations in vector space. In *ICLR*, 2013.

Ordonez, Vicente, Kulkarni, Girish, and Berg, Tamara L. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011.

Russakovsky, Olga, Deng, Jia, Su, Hao, Krause, Jonathan, Satheesh, Sanjeev, Ma, Sean, Huang, Zhiheng, Karpathy, Andrej, Khosla, Aditya, Bernstein, Michael S., Berg, Alexander C., and Fei-Fei, Li. Imagenet large scale visual recognition challenge. *IJCV*, 2015.

Silberman, Nathan, Hoiem, Derek, Kohli, Pushmeet, and Fergus, Rob. Indoor segmentation and support inference from RGBD images. In *ECCV*, 2012.

Simonyan, Karen and Zisserman, Andrew. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

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

Wu, Zhibiao and Palmer, Martha. Verb semantics and lexical selection. In *ACL*, 1994.

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

A. Supplementary Material
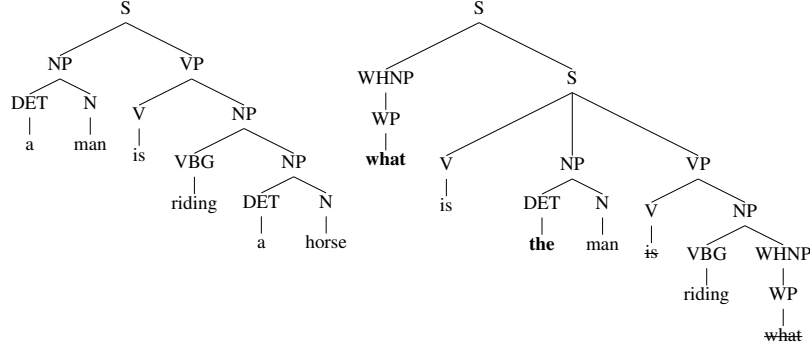
A.1. Question Generation: Syntax Tree Example

Figure 4. Example: "A man is riding a **horse**" $=>$ "**What** is **the** man riding?"

A.2. Post-Processing of COCO-QA Detail

First, answers that appear less than a frequency threshold are discarded. Second we enroll a QA pair one at a time. The probability of enrolling the next QA pair (q, a) is:

$$p(q, a) = \begin{cases} 1 & \text{if count}(a) \leq K \\ \exp\left(-\frac{count(a) - K}{[K_2}\right) & \text{otherwise} \end{cases} \tag{2}$$

where $count(a)$ denotes the current number of enrolled QA pairs that have a as the ground truth answer, and K, K_2 are some constants with $K \leq K_2$. In the COCO-QA generation we chose $K = 100$ and $K_2 = 200$.

A.3. More Sample Questions and Responses

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 11372
What do two women hold with a picture on it?
Ground truth: cake
IMG+BOW: cake (0.19)
2-VIS+BLSTM: cake (0.19)
BOW: umbrella (0.15)

DAQUAR 3018
What is on the right side?
Ground truth: table
IMG+BOW: tv (0.28)
2-VIS+LSTM: sofa (0.17)
LSTM: cabinet (0.22)

DAQUAR 1426
What is on the right side table?
Ground truth: tv
IMG+BOW: tv (0.25)
2-VIS+LSTM: tv (0.29)
LSTM: tv (0.14)

DAQUAR 1426a
What is on the left side of the room?
Ground truth: bed
IMG+BOW: door (0.19)
2-VIS+LSTM: door (0.25)
LSTM: door (0.13)

COCOQA 15756
What does the man rid while wearing a black wet suit?
Ground truth: surfboard
IMG+BOW: jacket (0.35)
2-VIS+LSTM: surfboard (0.53)
BOW: tie (0.30)

COCOQA 9715
What is displayed with the mattress off of it?
Ground truth: bed
IMG+BOW: bench (0.36)
2-VIS+LSTM: bed (0.18)
BOW: airplane (0.08)

COCOQA 25124
What is sitting in a sink in the rest room?
Ground truth: cat
IMG+BOW: toilet (0.77)
2-VIS+LSTM: toilet (0.90)
BOW: cat (0.83)

Figure 5. Sample questions and responses on Object questions

COCOQA 35
How many red velvet cup cakes with no frosting on a flowered plate?
Ground truth: three
IMG+BOW: three (0.43)
2-VIS+BLSTM: three (0.26)
BOW: two (0.26)

COCOQA 21446
How many different types of carrying cases in various colors?
Ground truth: four
IMG+BOW: four (0.69)
2-VIS+BLSTM: four (0.26)
BOW: three (0.25)

COCOQA 25600
How many double deckered busses parked near the green truck?
Ground truth: three
IMG+BOW: two (0.89)
2-VIS+BLSTM: two (0.44)
BOW: two (0.64)

DAQUAR 555
How many chairs are there?
Ground truth: one
IMG+BOW: six (0.20)
2-VIS+BLSTM: one (0.20)
LSTM: four (0.19)

*Figure 6.* Sample questions and responses on Number questions

COCOQA 298
What is the color of the airplane?
Ground truth: blue
IMG+BOW: blue (0.48)
2-VIS+LSTM: blue (0.33)
BOW: white (0.38)

DAQUAR 2989
What is the color of the sofa?
Ground truth: red
IMG+BOW: red (0.31)
VIS+LSTM-2: red (0.22)
LSTM: brown (0.21)

DAQUAR 2989a
What is the color of the table?
Ground truth: white
IMG+BOW: white (0.29)
VIS+LSTM-2: white (0.41)
LSTM: white (0.28)

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-2: yellow (0.26)
BOW: red (0.29)
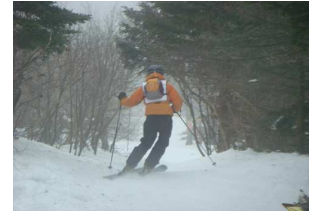
COCOQA 20140
What is the color of the coat?
Ground truth: orange
IMG+BOW: orange(0.21)
VIS+LSTM: yellow (0.21)
BOW: red (0.18)

COCOQA 20140a
What is the color of the pants?
Ground truth: black
IMG+BOW: black(0.27)
VIS+LSTM-2: black (0.23)
BOW: black (0.30)

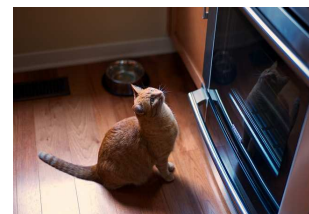*Figure 7.* Sample questions and responses on Color questions

COCOQA 6427
Where does the young boy secure himself?
Ground truth: suitcase
IMG+BOW: suitcase (0.39)
2-VIS+LSTM: car (0.32)
BOW: mirror (0.23)

COCOQA 6662
Where does the small kitten sit?
Ground truth: bowl
IMG+BOW: bowl (0.55)
2-VIS+LSTM: bowl (0.40)
BOW: chair (0.14)

COCOQA 28542
Where is the pan of food?
Ground truth: oven
IMG+BOW: oven (0.67)
2-VIS+LSTM: oven (0.93)
BOW: kitchen (0.37)

COCOQA 21112
Where is the cat observing the dishwasher?
Ground truth: kitchen
IMG+BOW: sink (0.39)
2-VIS+LSTM: chair (0.44)
BOW: sink (0.17)

Figure 8. Sample questions and responses on Location questions