# Ask Natural Questions About an Image
# Supplementary Material

## 1 Data Collection Prompt

The following instructions define what qualifies as VQG question, which served as our MTurk prompt:

INSTRUCTION .1

Imagine your friend shares with you an image without any specific prior context. Your task is to write "*the very first natural question*" that comes to your mind regarding the image you see. This question should be 'well-formed' and 'important' so that your friend would want to respond back. The answer to your question should not be obvious and evident in the image itself.

Following are important constraints that your questions should satisfy:
**1-** The answer to the question should not be obvious and evident in the image itself or obvious for anyone that can see the photo. You should ask a natural question that you would have asked your friend in a natural conversation, something that makes your friend want to respond back.

**2-** Do not ask generic questions that can be asked about "many other images". Ask questions which are really "specific to the image you see", and not any generic photo.

**3-** Each question should be a single question in one single sentence, grammatically correct with no spelling errors. Do not ask questions that have multiple parts. For example, do not ask 'Oh, that's cute, How old is the baby?', but ask 'How old is the baby?' instead.

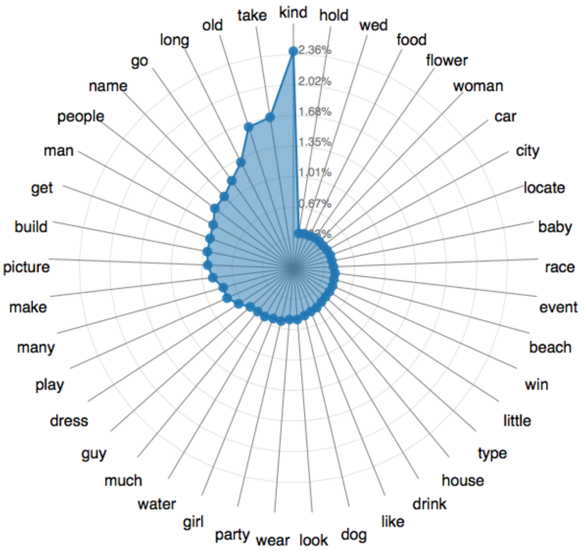


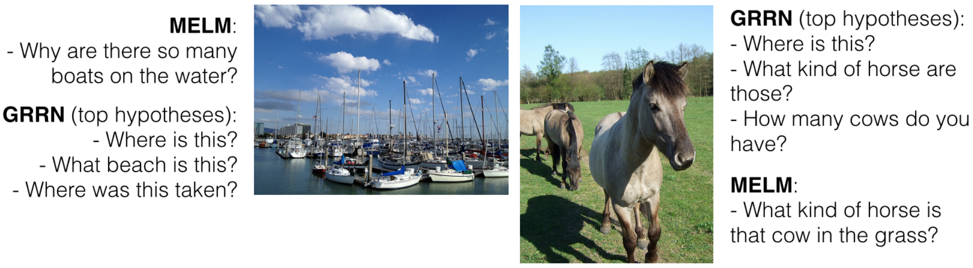
Figure 1: Frequency graph of top 40 words in $VQG_{Flickr-5000}$ dataset.

By instructing the workers to write questions which are exactly one complete sentence[1], we collect clean and to-the-point questions, which plays a crucial role in the quality of the resulting dataset.

## 2 $VQG_{Flickr-5000}$ Dataset

To collect more event-oriented images, we collected a new VQG dataset by sampling Flickr images that occur in non-initial and non-final position in a photo album; we hypothesized that albums are created for events. The frequency of words in $VQG_{Flickr-5000}$, as show in Figure 1, indicated that we still did not cover many of the commonsense events, such as *accident, earthquake* or *ceremony*, however, it does have a less biased range of concepts than MSCOCO. Figure 3 shows the radar graph of $VQG_{Bing-5000}$ dataset which meets our expectations.

[1] For example, a question such as 'Oh, that's cute, how old is the baby?' is not acceptable.



**MELM**:
- Why are there so many boats on the water?

**GRRN** (top hypotheses):
- Where is this?
- What beach is this?
- Where was this taken?

**GRRN** (top hypotheses):
- Where is this?
- What kind of horse are those?
- How many cows do you have?

**MELM**:
- What kind of horse is that cow in the grass?

Figure 2: Example generations on $VQG_{COCO-5000}$ using initial generation models.

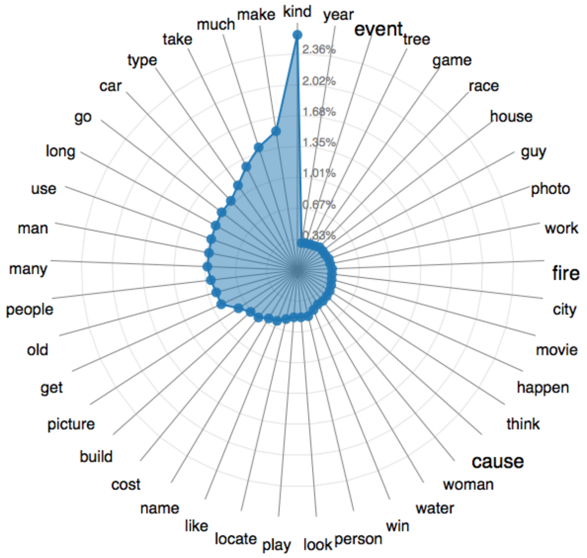


Figure 3: Frequency graph of top 40 words in $VQG_{Bing-5000}$ dataset.

## 3 Models

In this Section we provide some complementary details about the generative models for VQG.

### 3.1 MELM Model

We customized the state-of-the-art model in image captioning. This model (Fang et al., 2014), hereinafter MELM, is a pipeline starting from a set of candidate word probabilities (called MIL features) which is generated by a convolutional neural network (CNN) directly trained on images, then going through a maximum entropy (ME) language model which uses the MIL word probabilities to form a coherent sentence. We followed (Fang et al., 2014) and trained a MELM model on $VQG_{COCO-5000}$, with MIL features and fc7 features fine-tuned for detections on the MSCOCO dataset, to get the best visual accuracy. Figure 2 shows two questions generated by this model on two sample $VQG_{COCO-5000}$ test images. We observed that despite generating relevant questions content-wise, the MELM model made a few grammatical errors, improving which is a topic of further research and so we did not move forward with this model.

### 3.2 MT Model

This model showcases a text-to-text approach, where given a caption the system generates a questions. We constructed a parallel corpus of gold captions and VQG questions, using the gold captions of MSCOCO and $\boldsymbol{Captions_{Bing-5000}}$ dataset, total of 50,000 captions. The model consists of two recurrent neural networks (RNNs), an encoder which processes the caption and a decoder that generates the question. We specifically used a three-layer network with LSTM cells and attention mechanism. The trained network uses Stochastic Gradient Descent with early stopping, and decoded using a beam search of size 8. The outputs of this model tended to be very incoherent, often repeating a token multiple times. Given the size of our training data, we concluded that using captions as inputs to the network is not a promising approach.

### 3.3 GRNN

For the GRNN model described in the paper, while generating grammatical and coherent questions, 'where is this?' (Figure 2) was the chosen hypothesis for 29.3% of test images. This is a common problem with sequence-to-sequence models when they are trained towards optimizing log likelihood, where they prefer to generate the most generic sequence. In part, this behavior can be associated with relative higher frequency of generic questions in our dataset. However, if we look in the N-best list questions generated by this model, there are some meaningful and more interesting questions which rank lower. In order to address this issue, we added a constraint at decoding time to avoid generating questions which have less than 6 tokens. We use this GRNN model for our final experiments presented in the paper.

## 4 Human Evaluation

To ensure the quality of human evaluation, we set it up as follows: the human judge sees various system hypotheses at the same time in order to give a calibrated rating; moreover, the order of the system submissions is shuffled in order to ensure non-biased rating. Taking these issues into account, we crowdsourced our human evaluation on AMT. We tested different designs and prompts in pilot studies. Instruction .2 shows the final prompt we used in our evaluation framework.

INSTRUCTION .2

Imagine that your friend has shown you a photo. Your task is to rate each of the questions according to the following criterion. Your rating can be ‘disagree’, ‘in the middle’, or ‘agree’.

- This question is one of the first natural questions that comes to mind when my friend shows me this picture.

## References

Hao Fang, Saurabh Gupta, Forrest N. Iandola, Rupesh Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. 2014. From captions to visual concepts and back. *CoRR*, abs/1411.4952.