

# Image-Grounded Conversations: Multimodal Context for Natural Question and Response Generation

Nasrin Mostafazadeh<sup>1\*</sup>, Chris Brockett<sup>2</sup>, Bill Dolan<sup>2</sup>, Michel Galley<sup>2</sup>, Jianfeng Gao<sup>2</sup>,  
Georgios P. Spithourakis<sup>3\*</sup>, Lucy Vanderwende<sup>2</sup>

<sup>1</sup> University of Rochester, <sup>2</sup> Microsoft,

<sup>3</sup> University College London

nasrinm@cs.rochester.edu, chrisbkt@microsoft.com

## Abstract

The popularity of image sharing on social media reflects the important role visual context plays in everyday conversation. In this paper, we present a novel task, Image-Grounded Conversations (IGC), in which natural-sounding conversations are generated about shared photographic images. We investigate this task using training data derived from image-grounded conversations on social media and introduce a new dataset of crowd-sourced conversations for benchmarking progress. Experiments using deep neural network models trained on social media data show that the combination of visual and textual context can enhance the quality of generated conversational turns. In human evaluation, a gap between human performance and that of both neural and retrieval architectures suggests that IGC presents an interesting challenge for vision and language research.

## 1 Introduction

Significant advances in image captioning (Chen et al., 2015; Fang et al., 2014; Donahue et al., 2014; Chen et al., 2015) have enabled much interdisciplinary research in vision and language, from video transcription (Rohrbach et al., 2012; Venugopalan et al., 2015), to answering questions about images (Antol et al., 2015; Malinowski and Fritz, 2014), to storytelling around series of photos (Huang et al., 2016). Much of the focus has been on understanding images in terms of either describing (captioning)



**User1:** My son is ahead and surprised!

**User2:** Did he end up winning the race?

**User1:** Yes he won, he can't believe it!

Figure 1: A naturally-occurring Image-Grounded Conversation.

the image or answering questions about their content (Visual Question Answering (VQA)). In VQA, questions are constrained to be answerable from the image, i.e., they might be asked by someone unable to see the image. Understanding an image, however, involves more than captioning what is explicitly visible. Figure 1 illustrates a conversation between two users on social media. The conversation is grounded not only in visible objects (e.g., the boys, the bike) but more importantly, in events and actions (e.g., the race, winning) implied by the image. To humans viewing the images, these may be the most interesting and meaningful aspects. Visual Question Generation (VQG) (Mostafazadeh et al., 2016a) attempts to address the challenge of how to generate questions that involve such commonsense understanding of image content.

We extend VQG by introducing multimodal conversational context when formulating questions around images and training on naturally occurring social media data. To this end, we introduce the task of Image-Grounded Conversation (IGC), which requires a system to generate questions and responses in a natural-sounding conversation around a given image. IGC thus falls on the continuum between

\* This work was conducted at Microsoft.

chit-chat models and goal-directed conversation designed to accomplish a task. Visual grounding in an image constrains the topic while providing objects or inferrable events of interest so that a system can proactively drive the conversation forward. In this paper we focus principally on generating questions as conversation drivers.

This work thus draws together two threads of investigation that have hitherto remained largely unrelated: vision & language and data-driven conversation models. The contributions of this paper are threefold: (1) We extend VQG by introducing multimodal conversational context when formulating questions around images. To support this, we introduce the task of Image-Grounded Conversation (IGC) via a crowd-sourced dataset of 4,222 6-turn image-grounded conversations that will be publicly released, and compare IGC with other vision & language tasks by analyzing the characteristics of our IGC datasets and the effect of multimodal context (Section 4). (2) We investigate the application of deep neural generation and retrieval approaches for question and response generation tasks (Section 5), using models trained on 250K 3-turn image-grounded conversations found on Twitter and evaluated on our crowdsourced dataset. (3) Our experiments suggest that the combination of visual and textual context improves the quality of generated conversational turns and that visual context is more important than textual (Section 8). It is our hope that this work will furnish useful baselines to others working on multimodal conversation generation.

## 2 Related Work

### 2.1 Vision and Language

When trained on large datasets, such as the COCO dataset (Lin et al., 2014), Visual features combined with language modeling have shown good performance both in image captioning (Devlin et al., 2015; Fang et al., 2014; Donahue et al., 2014) and in Visual Question Answering (VQA) (Antol et al., 2015) and (Malinowski and Fritz, 2014). In VQA, questions are constrained to be answerable from the image, i.e., they might be asked by a person who cannot see the image. Das et al. (2016) extend the VQA scenario by collecting questions from people who are shown only an automatically generated caption, not the image itself, and demonstrate that system

performance is improved by treating questions as a series in a dialog rather than separate QA pairs. This form of dialog is best considered a simple one-sided QA exchange, in which only humans can ask questions and the system can only provide answers. (Ray et al., 2016) refine VQA by modeling whether the image contains enough information to answer the question; they observe that a model that can comment on the answerability of the question is preferable to a system that always answers.

Mostafazadeh et al. (2016a) introduce the task of visual question generation (VQG), in which the system itself outputs questions about the image. Questions are required to be ‘natural and engaging’, i.e. a person would find them interesting to answer, and may not be answerable from the image alone. In this work, we build on Mostafazadeh et al. (2016a) by introducing multimodal context when formulating questions and responses.

### 2.2 Data-Driven Conversational Modeling

This work is also closely linked to research on data-driven conversation modeling. Ritter et al. (2011) posed the response generation as a machine translation task, learning conversations from parallel message-response pairs found on social media. Their work has been successfully extended with the use of deep neural models (Sordani et al., 2015; Shang et al., 2015; Serban et al., 2015; Vinyals and Le, 2015; Li et al., 2016a; Li et al., 2016b). Sordani et al. (2015) introduce a context-sensitive neural language model that selects the most probable response conditioned on the conversation history (i.e., a text-only context). In this paper, we extend the contextual approach with the addition of multimodal features to build models that are capable of asking questions on topics of interests to a human, and that might allow a conversational agent to proactively drive the conversation forward.

## 3 Image-Grounded Conversations

For present purposes, we define the scope of IGC as the following two consecutive conversational steps:

- **Question Generation:** Given a visual context  $I$  and a textual context  $T$  (e.g., the first statement in Figure 1), generate a coherent, natural question  $Q$  about the image as the second utterance in the conversation. As seen in Figure 1, the question may not

be directly answerable from the image.

• **Response Generation:** Given a visual context  $I$ , a textual context  $T$ , and a question  $Q$ , generate a coherent, natural, response  $R$  to the question as the third utterance in the conversation. The response may be an answer, as expected in the VQA or Visual Dialog tasks, or it may be a comment, deflection, or other kind of response. The present work does not attempt time to generate responses from generated questions, a task that we leave to future work.

## 4 Data Collection

### 4.1 IGC<sub>Twitter</sub>

Previous work in neural conversation modeling (Ritter et al., 2010; Sordoni et al., 2015) has successfully used Twitter as the source of millions of natural conversations. In recent years, uploading a photo along with an accompanying tweet has become increasingly popular: multimedia tweets have risen 15% per year (as of June 2015, 28% total), with 42% of retweets containing non-verbal context (Morris et al., 2016). For training data, we sampled 250K quadruples of {visual context, textual context, question, response} tweet threads from a larger dataset of 1.4 million, extracted from the Twitter Firehose over a 3-year period beginning in May 2015 and filtered to select just those conversations in which the initial turn was associated with an image and the second turn was a question. Regular expressions were used to detect questions. To improve the likelihood that the authors are experienced Twitter conversationalists, we further limited extraction to those exchanges where users had actively engaged in at least 30 conversational exchanges during a 3-month period. Twitter data is notoriously noisy; we performed simple normalizations, and filtered out tweets that contained mid-tweet hashtags, were longer than 80 characters<sup>1</sup> and contained URLs not linking to the image. Table 1 presents example conversations from this dataset. Although the filters result in significantly higher quality of the extracted conversations, issues remain. A random sample of tweets suggests that about 46% of the Twitter conversations is affected by prior history between users, making response generation particularly difficult. In addition, the abundance of screen shots and non-

<sup>1</sup>Pilot studies showed that 80 character limit more effectively retains one-sentence utterances that are to the point.

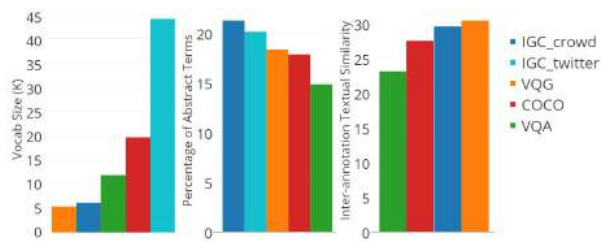


Figure 2: Comparison of V&L datasets.



Figure 3: Distribution of the number of tokens across datasets.

photograph graphics is potentially a major source of noise in extracting features for neural generation.

We use the IGC<sub>Twitter</sub> dataset as primary training data. For validation and test sets during model building, we held out a set of Twitter conversations, in which images and conversations had been vetted by crowd workers to be contentful and free of the kinds of image-related noise noted above.

### 4.2 IGC<sub>Crowd</sub>

To permit benchmarking of progress in the IGC task, we constructed test and validation datasets with more controlled parameters on the basis of the VQG dataset (Mostafazadeh et al., 2016a). We designed a crowdsourcing platform based on Turkserver (Mao et al., 2012), which enables synchronous and real-time interactions between crowd workers on Amazon Mechanical Turk (Mturk). Multiple workers wait in a virtual lobby to be paired with another worker who will be their conversation partner. After being paired, one of the users selects an image from a large photo gallery, after which the two users enter a chat window in which they have a short conversation about the selected image.

Images were sampled from the VQG dataset by querying a search engine using event-centric query terms that aggregated ‘event’ and ‘process’ hyponyms in WordNet (Miller, 1995) and using fre-





				
<b>Visual Context</b>				
<b>Textual Context</b>	Oh my gosh, i'm so buying this shirt.	I found a cawaii bird.	Stocking up!!	Only reason I come to carnival.
<b>Question</b>	Where did you see this for sale?	Are you going to collect some feathers?	Ayee! what the prices looking like?	Oh my God. How the hell do you even eat that?
<b>Response</b>	Midwest sports	There are so many crows here I'd be surprised if I never found one.	Only like 10-20% off..I think I'm gonna wait a little longer.	They are the greatest things ever chan. I could eat 5!

Table 1: Example conversations in the IGC<sub>Twitter</sub> dataset.

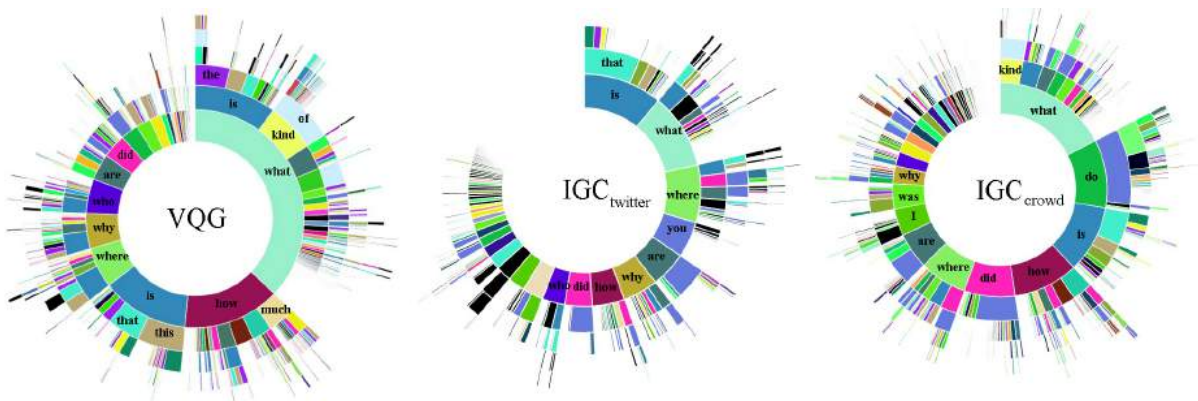


Figure 4: Distributions of n-gram sequences in questions in VQG, IGC<sub>Twitter</sub>, and IGC<sub>Crowd</sub>.

IGC <sub>Twitter</sub> (train set)	
# conversations = # images	<b>250k</b>
total # utterances	<b>750k</b>
IGC <sub>Twitter</sub> (val and test sets, split: 50% each)	
# conversations = # images	<b>4653</b>
total # utterances	<b>13,959</b>
IGC <sub>Crowd</sub> (val and test sets, split: 40% and 60%)	
# conversations = # images	<b>4,222</b>
total # utterances	<b>25,332</b>
average # utterances per conversation	4
# all workers participated	308
Max # conversations by one worker	20
Average work time per worker (min)	9.5
Median work time per worker (min)	10.0
IGC <sub>Crowd</sub> - <i>multiref</i> (val and test sets, split: 40% and 60%)	
# additional references per question/response	5
total # multi-reference utterances	<b>42,220</b>

Table 2: Basic Dataset Statistics.

quent TimeBank events (Pustejovsky et al., 2003). Table 3 shows three full conversations found in the IGC<sub>Crowd</sub> dataset. As the examples show, eventful images lead to conversations which are semantically very rich and would seem to require commonsense reasoning. Although the present work utilizes only three conversational turns, we collected up to six utterances per image for use in future work. To enable multi-reference evaluation (Section 6), we crowdsourced four additional questions and responses for the best IGC<sub>Crowd</sub> contexts and initial questions, as ranked by human annotators. The IGC<sub>Crowd</sub> dataset will be publicly released to the research community.

### 4.3 Dataset Characteristics

Table 2 summarizes basic dataset statistics. Figure 2 compares IGC questions with VQG and VQA questions in terms of vocabulary size, percentage of




Visual Context			
<b>Textual Context</b>	This wasn't the way I imagined my day starting.	I checked out the protest yesterday.	A terrible storm destroyed my house!
<b>Question</b>	do you think this happened on the highway?	Do you think America can ever overcome its racial divide?	OH NO, what are you going to do?
<b>Response</b>	Probably not, because I haven't driven anywhere except around town recently.	I can only hope so.	I will go live with my Dad until the insurance company sorts it out.
<b>Turn 4</b>	I would have to hate to change the tire on the highway.	Was the protest peaceful?	it's great that you can stay with someone!
<b>Turn 5</b>	Agreed, I should be grateful that I noticed it before I left my home.	Yes, it was, thankfully.	Yes he will be happy to have me for a while.
<b>Turn 6</b>	Call AAA, they will change it for you!	I hope the voices of minorities will be heard, and lead to changes in policing.	That's good to hear.
<b>VQG Question</b>	What caused that tire to go flat?	Where was the protest?	What caused the building to fall over?

Table 3: Example full conversations in our  $IGC_{Crowd}$  dataset. For comparison, we also include VQG questions in which the image is the only context.

abstract terms, and inter-annotation textual similarity. The COCO image captioning dataset is also included as a point of reference. The  $IGC_{Twitter}$  dataset has by far the largest vocabulary size, making it a more challenging dataset for training purposes. The  $IGC_{Crowd}$  and  $IGC_{Twitter}$ , in order, have the highest ratio of abstract to concrete terms. Broadly, abstract terms refer to intangibles, such as concepts, qualities, and feelings, whereas concrete terms refer to things that can be experienced with the five senses. It appears that conversational content may often involve abstract concepts than either captions or questions targeting visible image content.

It has been shown that humans achieve greater consensus on what a natural question to ask given an image (the task of VQG) than on captioning or asking a visually verifiable question (VQA) (Mostafazadeh et al., 2016b). The right-most plot in Figure 2 compares the inter-annotation textual similarity of our  $IGC_{Crowd}$  questions using a smoothed BLEU metric (Lin and Och, 2004).  $IGC_{Twitter}$  is excluded from this analysis as the data is not multireference. Contextually grounded questions of  $IGC_{Crowd}$  are competitive with VQG in inter-

annotation similarity. Figure 3 shows the distribution of the number of tokens per sentence. On average, the  $IGC_{Twitter}$  dataset has longer sentences. Figure 4 visualizes the n-gram distribution (with  $n=6$ ) of questions across datasets.  $IGC_{Twitter}$  is the most diverse set, with the lighter-colored part of the circle indicating sequences with less than 0.1% representation in the dataset.

**The Effectiveness of Multimodal Context:** The task of IGC emphasizes modeling of not only visual but also textual context. We presented human judges with a random sample of 600 triplets of image, textual context, and question ( $I, T, Q$ ) and asked them to rate the effectiveness of the image and the textual context, i.e., the degree to which the image or text is required in order for the sample question to sound natural. As Figure 5 shows, overall, both visual and textual contexts are indeed highly effective, and understanding both would be required for the question that was asked. We note that the crowd dataset more often requires understanding of the textual context than the Twitter set does.

**Frame Semantic Analysis:** The grounded conversations with questions in our datasets are full of

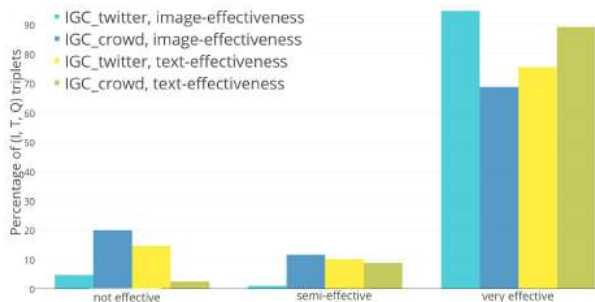


Figure 5: The effectiveness of textual and visual context for asking questions.

stereotypical commonsense knowledge. To get a better sense of the richness of our IGC<sub>Crowd</sub> dataset, we manually annotated a random sample of 330  $(I, T, Q)$  triplets in terms of Minsky’s Frames:<sup>2</sup> We annotated the FrameNet (Baker et al., 1998) frame evoked by the image ( $I_{FN}$ ), and then textual context ( $T_{FN}$ ). Then, for the asked question, we annotated the frame slot<sup>3</sup> ( $Q_{FN-slot}$ ) associated with a context frame ( $Q_{FN}$ ). These annotations can be accessed through <https://goo.gl/MVYGzP>. As the example in Table 4 shows, the image in isolation often does not evoke any uniquely contentful frame, whereas the textual context frequently does. In only 14% of cases does  $I_{FN}=T_{FN}$ , which further supports the complementary effect of our multimodal contexts. Moreover,  $Q_{FN}=I_{FN}$  for 32% our annotations, whereas  $Q_{FN}=T_{FN}$  for 47% of the triplets, again showing the effectiveness of textual context in determining the question to be asked.

## 5 Models

We use the VGGNet architecture (Simonyan and Zisserman, 2014) for computing deep convolutional image features. We primarily use the 4096-dimensional output of the last fully connected layer ( $fc7$ ) as the input to all the models sensitive to visual context.

<sup>2</sup>Minsky defines ‘frame’ as follows: “When one encounters a new situation, one selects from memory a structure called a Frame” (Minsky, 1974). According to Minsky, a frame is a commonsense knowledge representation data-structure for representing stereotypical situations, such as a wedding ceremony. Minsky further connects frames to the nature of questions: “[AFrame] is a collection of questions to be asked about a situation”. These questions can ask about the cause, intention, or side-effects of a presented situation.

<sup>3</sup>For 17% of cases we could not find a corresponding  $Q_{FN-slot}$  in FrameNet.


Visual Context	Textual Context	Question
	Look at all this food I ordered!	Where is that from?
FN Food	Request-Entity	Supplier

Table 4: FrameNet (FN) annotation of an example triplet.

## 5.1 Generation Models

Figure 6 overviews our three generation models. The conversation shown is based on the first conversation in Table 3.

**Visual Context Sensitive Model (V-Gen).** Similar to Recurrent Neural Network (RNN) models for image captioning (Devlin et al., 2015; Vinyals et al., 2015), (*V-Gen*) transforms the image feature vector to a 500-dimensional vector that serves as the initial recurrent state to a 500-dimensional one-layer Gated Recurrent Unit (GRU) which is the decoder module. The output sentence is generated one word at a time until the <EOS> (end-of-sentence) token is generated. We set the vocabulary size to 6000. Unknown words are mapped to an <UNK> token during training, which is not allowed to be generated at decoding time.

**Textual Context Sensitive Model (T-Gen).** This is a neural Machine Translation-like model that maps an input sequence to an output sequence (Seq2Seq model (Cho et al., 2014; Sutskever et al., 2014)) using an encoder and a decoder RNN. The decoder module is like the model described above, in this case the initial recurrent state being the 500-dimensional encoding of the textual context. For consistency, we use the same vocab size and number of layers as in the (*V-Gen*) model.

**Visual & Textual Context Sensitive Model (V&T-Gen).** This model fully leverages both textual and visual contexts. The vision feature is transformed to a 500-dimensional vector, and the textual context is likewise encoded into a 500-dimensional vector. The textual feature vector can be obtained using either a bag-of-words (*V&T.BOW-Gen*) representation, or an RNN (*V&T.RNN-Gen*), as depicted in Figure 7. The textual feature vector is then concatenated to the vision vector and fed into a fully connected (FC) feed forward neural network. As

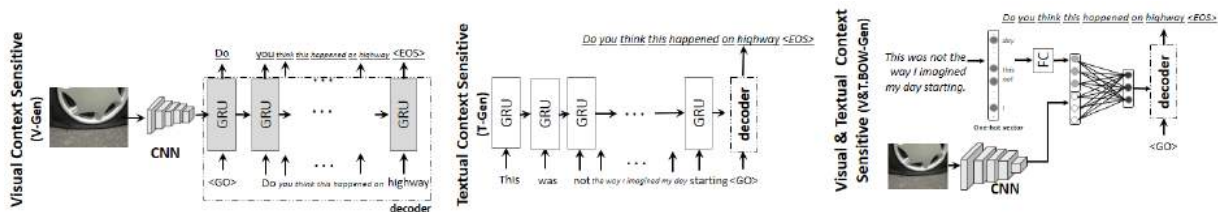


Figure 6: Question generation using the Visual Context Sensitive Model (*V-Gen*), Textual Context Sensitive Model (*T-Gen*), and the Visual & Textual Context Sensitive Model (*V&T.BOW-Gen*), respectively.

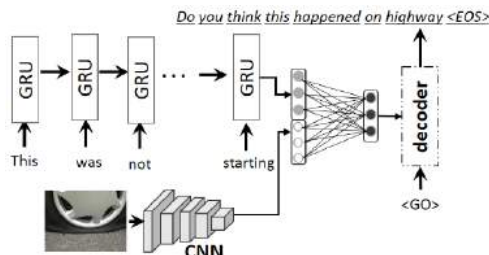


Figure 7: The visual & textual context sensitive model with RNN encoding (*V&T.RNN-Gen*).

a result, we obtain a single 500-dimensional vector encoding both visual and textual context, which then serves as the initial recurrent state of the decoder RNN.

In order to generate the response (the third utterance in the conversation), we need to represent the conversational turns in the textual context input. There are various ways to represent conversational history, including a bag of words model, or a concatenation of all textual utterances into one sentence (Sordani et al., 2015). For response generation, we implement a more complex treatment in which utterances are fed into an RNN one word at a time (Figure 7) following their temporal order in the conversation. An `<UTT>` marker designates the end of one utterance and the beginning of the next.

**Decoding and Reranking.** For all generation models, at decoding time we generate the N-best lists using left-to-right beam search with *beam-size* 25. We set the maximum number of tokens to 13 for the generated partial hypotheses. Any partial hypothesis that reaches `<EOS>` token becomes a viable full hypothesis for reranking. The first few hypotheses on top of the N-best lists generated by Seq2Seq models tend to be very generic,<sup>4</sup> disregarding the input context. In order to address this issue we rerank

<sup>4</sup>An example generic question is *where is this?* and a generic response is *I don't know*.

the N-best list using the following score function:

$$\log p(h|C) + \lambda \text{idf}(h, D) + \mu|h| + \kappa V(h) \quad (1)$$

where  $p(h|C)$  is the probability of the generated hypothesis  $h$  given the context  $C$ . The function  $V$  counts the number of verb POS in the hypothesis and  $|h|$  denotes the number of tokens in the hypothesis. The function  $\text{idf}$  is the inverse document frequency, simply computing how common a hypothesis is across all the generated N-best lists. Here  $D$  is the set of all N-best lists and  $d$  is a specific N-best list. We define  $\text{idf}(h, D) = \log \frac{|D|}{|\{d \in D: h \in d\}|}$ , where we set  $N=10$  to cut short each N-best list. We optimize all the parameters of the scoring function towards maximizing the smoothed-BLEU score (Lin and Och, 2004) using the Pairwise Ranking Optimization algorithm (Hopkins and May, 2011).

## 5.2 Retrieval Models

In addition to generation, we implemented two retrieval models customized for the tasks of question and response generation. Work in vision and language has demonstrated the effectiveness of retrieval models, where one uses the annotation (e.g., caption) of a nearest neighbor in the training image set to annotate a given test image (Mostafazadeh et al., 2016a; Devlin et al., 2015; Hodosh et al., 2013; Ordonez et al., 2011; Farhadi et al., 2010).

**Visual Context Sensitive Model (*V-Ret*).** This model only uses the given image for retrieval. First, we find a set of  $K$  nearest training images for the given test image based on cosine similarity of the *fc7* vision feature vectors. Then we retrieve those  $K$  annotations as our pool of  $K$  candidates. Finally, we compute the textual similarity among the questions in the pool according to a Smoothed-BLEU (Lin and Och, 2004) similarity score, then emit the sentence with the highest similarity to the rest of the pool.




	Visual Context			
Question Generation	<b>Textual Context</b>	The weather was amazing at this baseball game.	I got in a car wreck today!	My cousins at the family reunion.
	<b>Gold Question</b>	Nice, which team won?	Did you get hurt?	What is the name of your cousin in the blue shirt?
	<b>V&amp;T-Ret</b>	U at the game? or did someone take that pic for you?	<b>You driving that today?</b>	<b>U had fun?</b>
	<b>V-Gen</b> <b>V&amp;T-Gen</b>	Where are you? <b>Who's winning?</b>	Who's is that? <b>What happened?</b>	Who's that guy? <b>Where's my invite?</b>
Response Generation	<b>Textual Context</b>	The weather was amazing at this baseball game. <UTT> Nice, which team won?	I got in a car wreck today! <UTT> Did you get hurt?	My cousins at the family reunion. <UTT> What is the name of your cousin in the blue shirt?
	<b>Gold Response</b>	My team won this game.	No it wasn't too bad of a bang up.	His name is Eric.
	<b>V&amp;T-Ret</b>	10 for me and 28 for my dad.	<b>Yes.</b>	lords cricket ground . beautiful.
	<b>V&amp;T-Gen</b>	ding ding ding!	<b>Nah, I'm at home now.</b>	<b>He's not mine!</b>

Table 5: Example baseline question and response generations on  $IGC_{Crowd}$  test set. All the generation models use beam search with reranking. In the textual context, <UTT> separates different utterances. The generations in bold are acceptable utterances given the underlying context.

**Visual & Textual Context Sensitive Model (V&T-Ret).** This model uses both visual and textual contexts to retrieve a question or a response. A linear combination of  $fc7$  and word2vec feature vectors is utilized for retrieving similar training instances.

## 6 Evaluation Setup

We provide both human and automatic evaluations for our question and response generation tasks. We crowdsource our human evaluation on an AMT-like crowdsourcing system, asking seven crowd workers to each rate the quality of candidate questions or responses on a three-point Likert-like scale, ranging from 1 to 3 (the highest). In order to ensure a calibrated rating, we show the human judges all system hypotheses for a particular test case at the same time. System outputs were randomly ordered to prevent judges from guessing which systems were which on the basis of position. After collecting judgments, we averaged the scores throughout the test set for each model. We discarded as spammers all annotators whose ratings varied from the mean by more than 2 standard deviations.

Although human evaluation is to be preferred, and currently essential, in open-domain generation tasks involving intrinsically diverse outputs, it is useful to have an automatic metric for day-to-day evaluation. For ease of replicability, we use the standard Machine Translation metric, BLEU (Papineni et al., 2002), which captures n-gram overlap between hypotheses and references. Results reported below employ BLEU with equal weights up to 4-grams.

## 7 Experimental Results

We experiment with all the models presented in Section 5. For question generation, we use a visual & textual sensitive model that uses bag-of-words ( $V\&T.BOW-Gen$ ) to represent the textual context, which achieved better results. Earlier vision & language work such as VQA (Antol et al., 2015) has shown that a bag-of-words baseline outperforms LSTM-based models for representing textual input when visual features are available (Zhou et al., 2015). In response generation, which needs to account for textual input consisting of two turns, we use the  $V\&T.RNN-Gen$  model as the visual &



	Human	Generation (Greedy)			Generation (Beam, best)				Generation (Reranked, best)			Retrieval		
		Gold	Textual	Visual	V & T	Textual	Visual	V & T	VQG	Textual	Visual	V & T	Visual	V & T
<i>Q.</i>	Twitter	2.26	1.39	2.13	1.57	1.06	<b>2.35</b>	1.90	1.49	1.03	1.72	1.71	1.71	1.68
	Crowd	<u>2.68</u>	1.46	1.58	1.86	1.07	1.86	<b>2.28</b>	2.24	1.03	2.06	2.13	1.59	1.54
<i>R.</i>	Twitter	<u>2.44</u>	1.26	–	1.60	1.13	–	<b>1.68</b>	–	1.05	–	1.60	–	1.59
	Crowd	<u>2.75</u>	1.24	–	1.40	1.12	–	<b>1.49</b>	–	1.04	–	1.44	–	1.48

Table 6: Human judgment results. The maximum score is 3. The *Q.* rows correspond to the question generation task and the *R.* rows correspond to response generation. Per model, the human score is computed by averaging across multiple images. The boldfaced numbers show the highest score among the systems and underline signifies the overall highest score.

	Textual	Generation			Retrieval	
		Visual	V & T	VQG	Visual	V & T
<i>Q.</i>	Twitter	0.94	2.06	<b>2.13</b>	0.37	0.75
	Crowd	0.65	0.9	<b>1.25</b>	1.23	0.44
	Crowd <sub>m</sub>	1.71	3.23	4.41	<b>8.61</b>	0.76
<i>R.</i>	Twitter	0.35	–	<b>0.44</b>	–	0.24
	Crowd	0.0	–	<b>0.29</b>	–	0.15
	Crowd <sub>m</sub>	1.34	–	<b>1.57</b>	–	0.66

Table 7: Results of evaluating various models according to automatic metric. Crowd<sub>m</sub> refers to the multi-reference test set.

textual-sensitive model in the *R.* rows of tables 6 and 7. Since generating a response solely from the visual context is unlikely to be successful, we do not use the *V-Gen* model in response generation.

Table 5 presents a few example generations by our best performing systems. Tables 6 and 7 provide the human and automatic evaluation results for all of our models. All models are trained on IGC<sub>Twitter</sub> (training set), except for the model labeled VQG, which is the same (*V-Gen*) model, but trained on 7,500 questions from the VQG dataset (Mostafazadeh et al., 2016b). All systems have been tuned/tested on the corresponding IGC dataset. As a point of reference, we include the gold standard human reference in the human evaluations.

In human evaluation, the model that encodes both visual and textual context outperforms others, except for the question generation on Twitter in which the visual model wins. It appears the visual context is more effective in Twitter dataset, as we have shown in Section 4. We note that human judges preferred the top generation in the n-best list over the reranked best, likely due to tradeoff between a safe and generic utterance and a riskier but contentful one. As shown in Table 6, our best performing

generation system scores higher than human on the Twitter test set for question generation, but otherwise the human gold references in our Crowd set are consistently favored throughout the table. We take this as evidence that IGC<sub>Crowd</sub> provides a robust and challenging test set for benchmarking the progress on the task.

BLEU scores are low, as is characteristic for language tasks with intrinsically diverse outputs (Li et al., 2016b; Li et al., 2016a). Automatic evaluation in Table 7 provides confirmation that the IGC<sub>Crowd</sub> test set is more challenging to models that have been trained on IGC<sub>Twitter</sub>. On BLEU, the multimodal *V&T* model outperforms all the other models across test sets, except for the multireference test set (Crowd<sub>m</sub> in Table 7), in which the VQG model does significantly better. We attribute this to differences in the Twitter and our Crowd datasets, as discussed in Section 4.

Overall, in both automatic and human evaluation, our question generation models are more successful than response generation. More sophisticated models and larger training data sets may overcome this disparity in the future.

## 8 Conclusions

We have introduced a new task of multimodal image-grounded conversation, in which, when given an image and a natural language text, the system must generate meaningful conversation turns. To support this task, we are releasing to the research community a crowdsourced dataset of 4,222 high-quality conversations about eventful images with up to 6 turns each, and multiple references.

Our experiments provide evidence that capturing multimodal context improves the quality of generation. The gap between the performances of our best

models and humans opens opportunities further research in the continuum from casual conversation to more task- and topic-oriented vision and language dialog. We expect also that addition of other kinds of grounding may further improve performance of systems.

## References

- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *International Conference on Computer Vision (ICCV)*.
- Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *Proceedings of the 17th International Conference on Computational Linguistics - Volume 1*, COLING '98, pages 86–90, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Jianfu Chen, Polina Kuznetsova, David Warren, and Yejin Choi. 2015. Déjà image-captions: A corpus of expressive descriptions in repetition. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 504–514, Denver, Colorado, May–June. Association for Computational Linguistics.
- Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.
- Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2016. Visual dialog. *CoRR*, abs/1611.08669.
- Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. 2015. Language models for image captioning: The quirks and what works. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 100–105, Beijing, China, July. Association for Computational Linguistics.
- Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2014. Long-term recurrent convolutional networks for visual recognition and description. *CoRR*, abs/1411.4389.
- 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.
- Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *Proceedings of the 11th European Conference on Computer Vision: Part IV, ECCV'10*, pages 15–29, Berlin, Heidelberg. Springer-Verlag.
- Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Int. Res.*, 47(1):853–899, May.
- Mark Hopkins and Jonathan May. 2011. Tuning as ranking. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1352–1362, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.
- Ting-Hao (Kenneth) Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239, San Diego, California, June. Association for Computational Linguistics.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California, June. Association for Computational Linguistics.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.
- Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics, ACL '04*, Strouds-

- burg, PA, USA. Association for Computational Linguistics.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick, 2014. *Microsoft COCO: Common Objects in Context*, pages 740–755. Springer International Publishing, Cham.
- Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in Neural Information Processing Systems 27*, pages 1682–1690.
- Andrew Mao, David Parkes, Yiling Chen, Ariel D. Procaccia, Krzysztof Z. Gajos, and Haoqi Zhang. 2012. Turkserver: Enabling synchronous and longitudinal online experiments. In *Workshop on Human Computation (HCOMP)*.
- George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November.
- Marvin Minsky. 1974. A framework for representing knowledge. Technical report, Cambridge, MA, USA.
- Meredith Ringel Morris, Annuska Zolyomi, Catherine Yao, Sina Bahram, Jeffrey P. Bigham, and Shaun K. Kane. 2016. "with most of it being pictures now, I rarely use it": Understanding twitter's evolving accessibility to blind users. In Jofish Kaye, Allison Druin, Cliff Lampe, Dan Morris, and Juan Pablo Hourcade, editors, *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, May 7-12, 2016*, pages 5506–5516. ACM.
- Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016a. Generating natural questions about an image. In *Proceedings of the Annual Meeting on Association for Computational Linguistics, ACL '16*. Association for Computational Linguistics.
- Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016b. Generating natural questions about an image. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1802–1813, Berlin, Germany, August. Association for Computational Linguistics.
- Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Neural Information Processing Systems (NIPS)*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02*, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.
- J. Pustejovsky, P. Hanks, R. Sauri, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro, and M. Lazo. 2003. The TIMEBANK corpus. In *Proceedings of Corpus Linguistics 2003*, pages 647–656, Lancaster, March.
- Arijit Ray, Gordon Christie, Mohit Bansal, Dhruv Batra, and Devi Parikh. 2016. Question relevance in VQA: identifying non-visual and false-premise questions. In Jian Su, Xavier Carreras, and Kevin Duh, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 919–924. The Association for Computational Linguistics.
- Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Un-supervised modeling of twitter conversations. In *In HLT-NAACL*.
- Alan Ritter, Colin Cherry, and William B Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 583–593. Association for Computational Linguistics.
- Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. 2012. A database for fine grained activity detection of cooking activities. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, IEEE, June.
- Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2015. Hierarchical neural network generative models for movie dialogues. *arXiv preprint arXiv:1507.04808*.
- Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *ACL-IJCNLP*.
- K. Simonyan and A. Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.
- Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205, Denver, Colorado, May–June. Association for Computational Linguistics.
- Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.
- Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate

- Saenko. 2015. Translating videos to natural language using deep recurrent neural networks. In *Proceedings the 2015 Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL HLT 2015)*, pages 1494–1504, Denver, Colorado, June.
- Oriol Vinyals and Quoc Le. 2015. A neural conversational model. In *Deep Learning Workshop, ICML*.
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition*.
- Bolei Zhou, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. 2015. Simple baseline for visual question answering. *CoRR*, abs/1512.02167.