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

Mostafazadeh et al. (2017)
University of Rochester, Microsoft and University College London

Presenter: Shubham Annadate

Agenda

- Introduction
- Related Work
- Image Grounded Conversation (IGC)
  - Dataset
  - Task Characteristics
- Models
- Evaluation Setup
- Experimental Results
- Conclusion/Key Contributions
- Thoughts

Introduction

- Recent work on vision and language → describing or answering questions about image.
- Conversations threads on social media platforms like Twitter.
  - 28% of tweets contain image (as of 2015)
  - Conversation are beyond what is explicitly visible in the image

User1: My son is ahead and surprised!

User2: Did he end up winning the race?

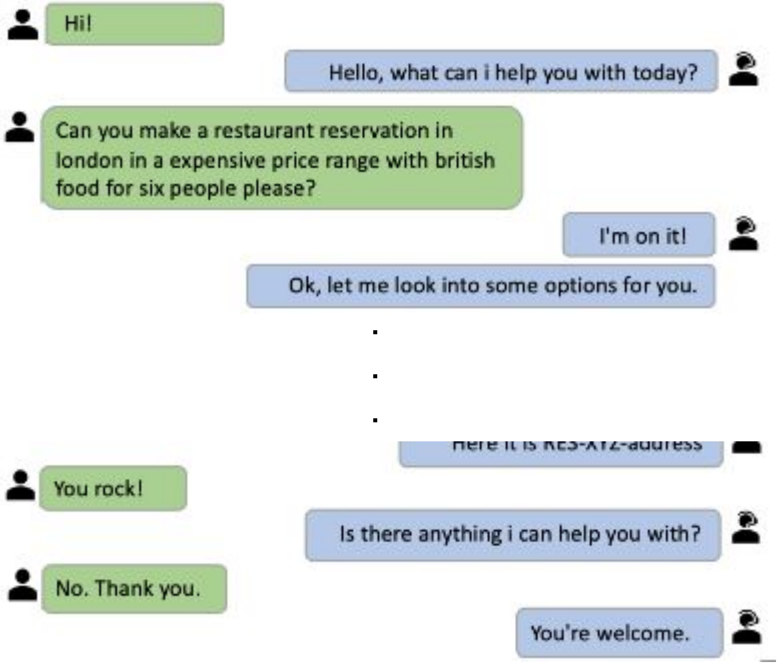
User1: Yes he won, he can't believe it!

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

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

Introduction

- Look at image as a context for interaction rather than an artifact
- Image-Grounded Conversation (IGC)
    - System that generates conversational turns to drive conversation.
    - Falls between chit-chat and goal-oriented dialog systems
    - Combines the threads of language & vision and data-driven conversational modeling



Goal-oriented dialog systems

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

Related Work

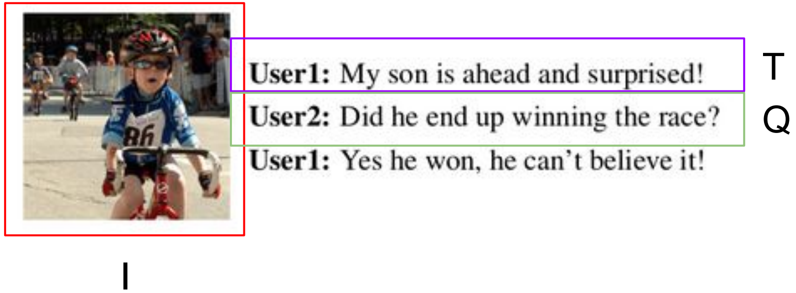
- Vision and Language
  - Visual features combined with language modeling have shown improved performance in image captioning and VQA tasks (2014-15)
  - **Visual Question Generation (VQG) Task** (2016)
- Data-Driven Conversational Modeling
  - Learning conversations from message-response pairs (2011)
  - Context-Sensitive Neural Language Models (2015-16)

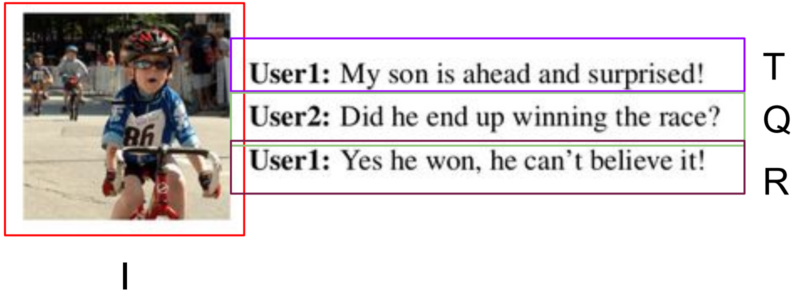
IGC (Task Definition)

Two consecutive conversational steps within the current scope:

Question generation: *(I, T)* → *Q*

Response Generation: *(I, T, Q)* → *R*





I: visual context, T: textual context, Q: question, R: response

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

IGC (Dataset)

- No pre-existing dataset for IGC task
- IGC_{Crowd}
  - Sampled *eventful* images from VQG dataset
  - Pair of Amazon MTurk workers have a short conversation about the image
  - For multi-reference evaluation → crowd-sourced 5 additional references per question/response.
  - Used for **validation and testing purpose**
- $IGC_{Twitter}$
  - Used for **training purpose**
  - Sampled 250K quadruples of {I, T, Q, R} tweet threads from Twitter dataset

IGC_{Crowd} (val and test sets, split: 40% and 60%)	
# conversations = # images	**4,222**
total # utterances	**25,332**
# all workers participated	308
Max # conversations by one worker	20
Average payment per worker (min)	1.8 dollars
Median work time per worker (min)	10.0
$IGC_{Crowd-multiref}$ (val and test sets, split: 40% and 60%)	
# additional references per question/response	5
total # multi-reference utterances	**42,220**

Table 2: Basic Dataset Statistics.

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

IGC$_{\text{Crowd}}$ (Dataset)

Visual Context			
Textual Context	This wasn't the way I imagined my day starting.	I checked out the protest yesterday.	A terrible storm destroyed my house!
Question	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?
Response	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.
VQG Question	What caused that tire to go flat?	Where was the protest?	What caused the building to fall over?

Table 1: Example full conversations in our IGC$_{\text{Crowd}}$ dataset. For comparison, we also include VQG questions in which the image is the only context.

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

Task Characteristics (Effectiveness of Multimodal Context)

- IGC task emphasizes modeling both visual and textual context
- Presented human judges with 600 (I,T,Q) triplets from each $IGC_{Twitter}$ and IGC_{Crowd} and asked to rate *effectiveness* of visual and textual context
  - Whether the question makes sense without the image or text?

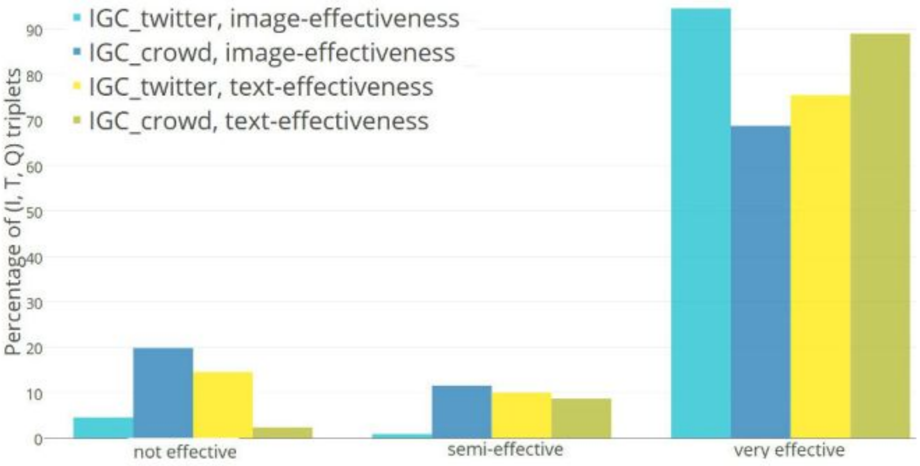


Figure 3: The effectiveness of textual and visual context for asking questions.

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

Task Characteristics (Frame Semantic Analysis of Questions)

- Manually annotate a sample of 330 (I,T,Q) triplets in terms of Minsky's Frames
  - Frame: semantic representation of a situation involving participants, props and other conceptual roles
- Annotated the FrameNet frame evoked by I, T and Q.
  - 14% $I_{FN} = T_{FN}$
  - 32% $Q_{FN} = I_{FN}$
  - 47% $Q_{FN} = T_{FN}$

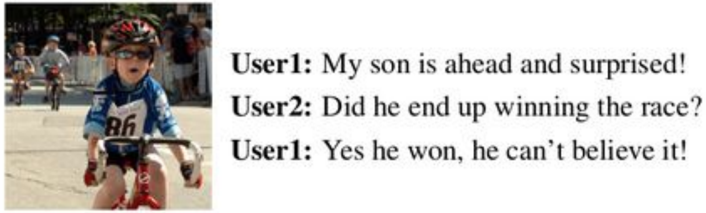
	Visual Context	**Textual Context**	**Question**
		Look at all this food I ordered!	Where is that from?
FN	*Food*	*Request-Entity*	*Supplier*

Table 3: FrameNet (FN) annotation of an example.

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

Task Characteristics (Event Analysis of Conversations)

Manually annotated 20 conversations with their causal and temporal event structure (CaTeRS Scheme)



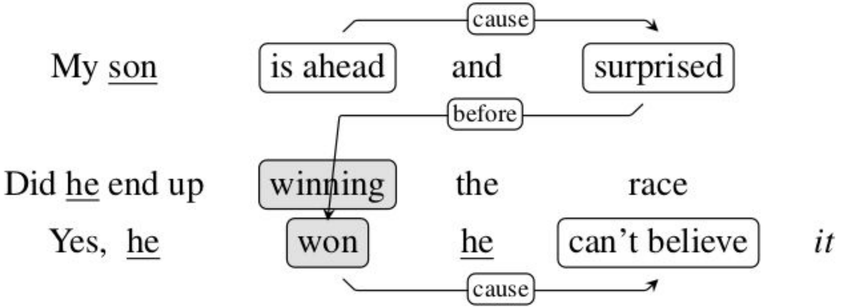


Figure 4: An example causal and temporal (CaTeRS) annotation on the conversation pre-

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

Task Characteristics (Event Analysis of Conversations)

- Findings:
  - IGC utterances are rich in events (0.71 event entity mentions)
  - Semantic link annotations reflect common sense relations between event mentions in context of ongoing conversation
  - **Capturing causal and temporal relations between events is necessary for a system to successfully perform IGC task**

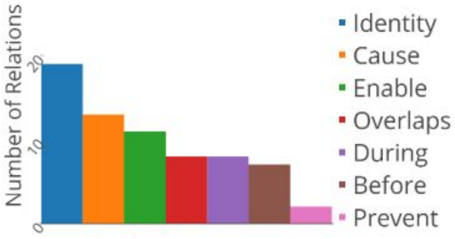


Figure 5: The frequency of event-event semantic links in a random sample of 20 IGC conversations.

Models (Question Generation Models) *(I, T) → Q*

1. Visual Context Sensitive Model (V-Gen)

Visual Context	
Textual Context	This wasn't the way I imagined my day starting.
Question	do you think this happened on the highway?
Response	Probably not, because I haven't driven anywhere except around town recently.

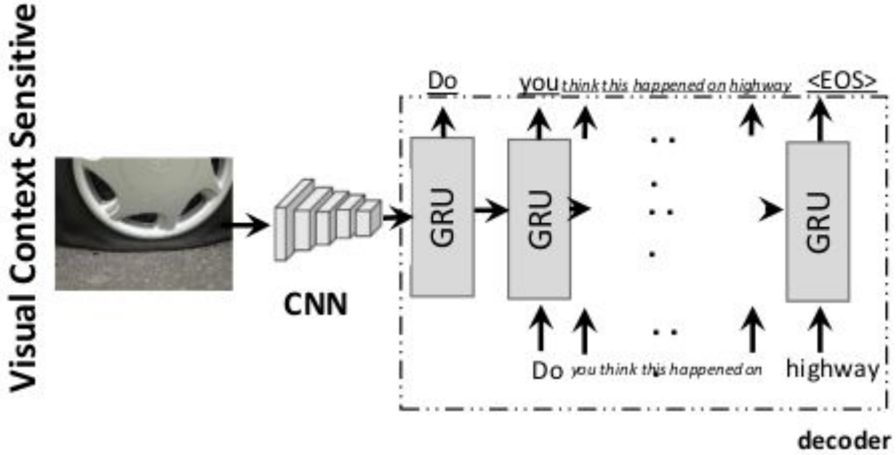


Image-Grounded Conversations: Multimodal Context for Natural Question and Response Generation

Models (Question Generation Models)   *(I, T) → Q*

2. Textual Context Sensitive Model (T-Gen)

Visual Context	
Textual Context	This wasn't the way I imagined my day starting.
Question	do you think this happened on the highway?
Response	Probably not, because I haven't driven anywhere except around town recently.

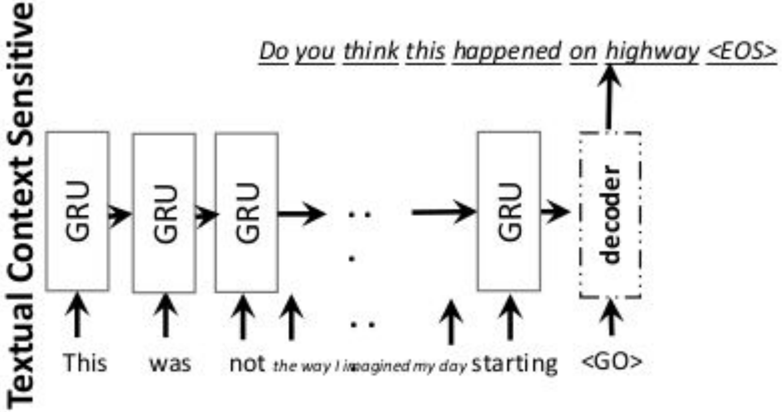
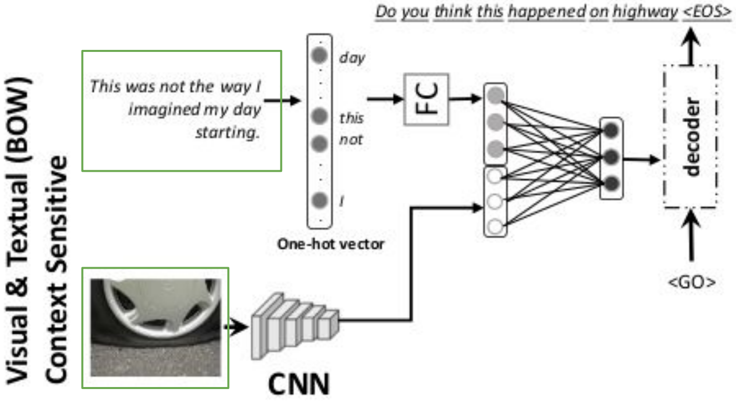


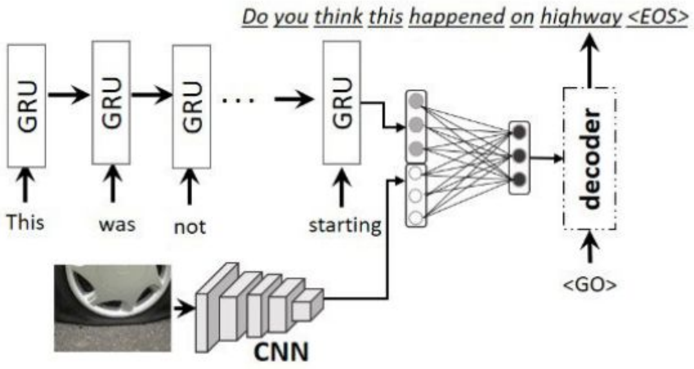
Image-Grounded Conversations: Multimodal Context for Natural Question and Response Generation

Models (Question Generation Models) *(I, T) → Q*

3. Visual & Textual Context Sensitive Model (V&T-Gen)



V&T.BOW-Gen



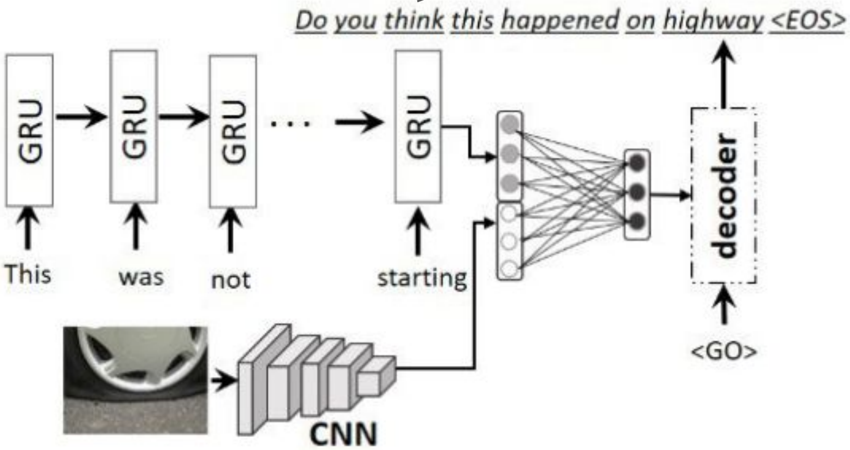
V&T.RNN-Gen

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

Models (Response Generation Models) *(I, T, Q) → R*

Visual Context	
Textual Context	This wasn't the way I imagined my day starting.
Question	do you think this happened on the highway?
Response	Probably not, because I haven't driven anywhere except around town recently.

Probably not…. recently <EOS>



Textual context:
This was not the way I imagined my day starting ***<UTT>*** *Do you think this happened on the highway?*

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

Models (Decoding and Reranking)

applies to generation models

- Greedy Decoding
- Beam Search
  - Generate N-best lists using left-to-right beam search (beam size = 25)
  - Max #tokens = 13
  - Any partial hypothesis that reaches <EOS> → viable for reranking
- Reranking
  - First few hypotheses on top of the N-best list tend to be generic
    - For example, *"Where is this?"*

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

Probability of hypothesis(h) given context (C)

How common the hypothesis is across all generated N-best lists (D)

tokens in hypothesis

verb POS in hypothesis

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

Models (Retrieval Models)

1. Visual Context Sensitive Model (V-Ret)
    a. Only uses the given image for retrieval
    b. Finds K nearest training images for the given test image based on cosine similarity of fc7 feature vector → K candidates
    c. Compute textual similarity among the questions in the pool (Smoothed BLEU similarity score)
    d. Emit sentence with highest similarity to rest of the pool.
2. Visual & Textual Context Sensitive Model (V&T-Ret)
    a. Uses linear combination of fc7 and word2vec feature vectors
    b. Retrieval process is same as above

Models (Recap)

- Question/Response generation
  - V-Gen
  - T-Gen
  - V&T-Gen
    - V&T.BOW-Gen (question)
    - V&T.RNN-Gen (response)
- Retrieval
  - V-Ret
  - V&T Ret
- Train: $IGC_{Twitter}$
- Validation/Test: IGC_{Crowd}

- greedy
- beam search (best)
- reranked (best)

Evaluation Setup

- Human evaluation
  - Crowdsource on AMT-like system
  - 7 crowd workers rate quality of questions/responses on a scale of 1 to 3 (highest)
  - All system hypotheses are presented at the same time in random order
  - Also present the human gold standard
  - Average the score throughout the test set for each model and the human gold standard
- Automatic Evaluation
  - BLEU with equal weight upto 4 grams at corpus level on the multi-reference IGC_{Crowd} test set

Experimental Results (Human Evaluation)

	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
Question	2.68	1.46	1.58	1.86	1.07	1.86	**2.28**	2.24	1.03	2.06	2.13	1.59	1.54
Response	2.75	1.24	–	1.40	1.12	–	**1.49**	–	1.04	–	1.44	–	1.48

Key Takeaways:

- Multimodal V&T outperforms Textual and Visual
- Top generation in N-best list is preferred over reranked
- Human gold standards are favoured throughout the table
- $\text{IGC}_{\text{Crowd Test Set}}$ robust test set for benchmarking progress on this task

Experimental Results (Automatic Evaluation)

	Textual	Generation			Retrieval	
		Visual	V & T	VQG	Visual	V & T
Question	1.71	3.23	4.41	**8.61**	0.76	1.16
Response	1.34	–	**1.57**	–	–	0.66

Key Takeaways:

- BLEU scores are low
- Multimodal V&T outperforms all other models except VQG
- For both automatic and human evaluation → performance on question generation is better than response generation

Experimental Results (Examples)

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

Table 4: Example question and response generations on $\text{IGC}_{\text{Crowd}}$ 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.

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

Conclusion/Contributions

- Introduced a new task on multimodal image-grounded conversation for formulating questions and responses around images.
- Released to research community a crowdsourced dataset of 4222 high quality multi-turn conversations around eventful images and multiple references.
- Experiments suggest that capturing multimodal context improves the quality of question and response generation

Thoughts

- First step into combining threads of language & vision and conversations
- Including other kinds of grounding
  - Temporal, geolocation …
- Attention based mechanism
- Performance across different automatic evaluation metrics
- More to a conversation than just question and response
- Better quality training dataset compared to $IGC_{Twitter}$

IGC$_{\text{Twitter}}$ Training Dataset Problems

- Conversation is not grounded in image and textual context but rather in participant's established relation or prior history
  - Around 46% are like this!
- Abundance of screenshots

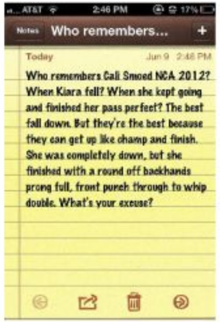
	
Smile. Why are you so obsessed with me? Oh pls	What's your excuse? Nca nationals? which day? Day 2 i believe ! if you go on youtube it should be the first one !

Table 8: Example Twitter conversations that add noise to the dataset.

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

Questions/Comments?

Supplementary Material

IGC$_{\text{Twitter}}$ Training Dataset Example

Visual Context				
Textual Context	Oh my gosh, i'm so buying this shirt.	I found a cawaii bird.	Stocking up!!	Only reason I come to carnival.
Question	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?
Response	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 7: Example conversations in the IGC$_{\text{Twitter}}$ dataset.

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

IGC$_{\text{Twitter}}$ Training Dataset Problems

- Conversation is not grounded in image and textual context but rather in participant's established relation or prior history
  - Around 46% are like this!
- Abundance of screenshots

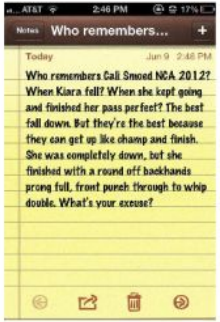
	
Smile. Why are you so obsessed with me? Oh pls	What's your excuse? Nca nationals? which day? Day 2 i believe ! if you go on youtube it should be the first one !

Table 8: Example Twitter conversations that add noise to the dataset.

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

Data Analysis (length of sentences)

On average, $IGC_{Twitter}$ has longer sentences

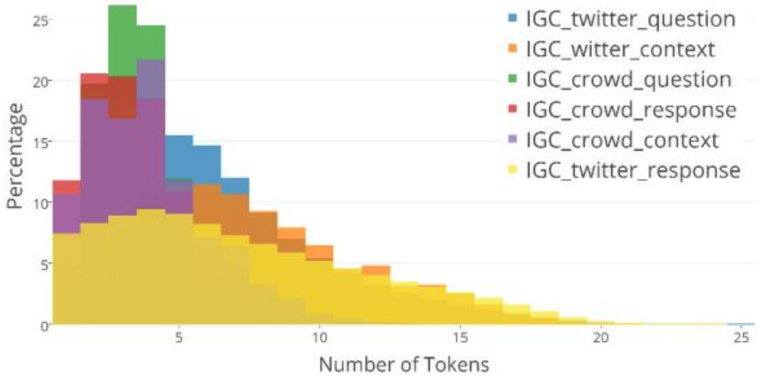


Figure 8: Distribution of the number of tokens across datasets.

Data Analysis (diversity in questions)

Figure 9: Distributions of n-gram sequences in questions in VQG, $IGC_{Twitter}$, and IGC_{Crowd}.

$IGC_{Twitter}$ is most diverse, with light-colored part of the circle indicating sequences with less than 0.1% representation in the data.

Data Analysis (IGC questions characteristics)

- IGC$_{\text{Twitter}}$ has largest vocabulary size → challenging for training
- IGC$_{\text{Twitter}}$ and IGC$_{\text{Crowd}}$ have highest ratio of concrete to abstract nouns
- Contextually grounded questions of IGC$_{\text{Crowd}}$ are competitive with VQG in inter-annotation similarity

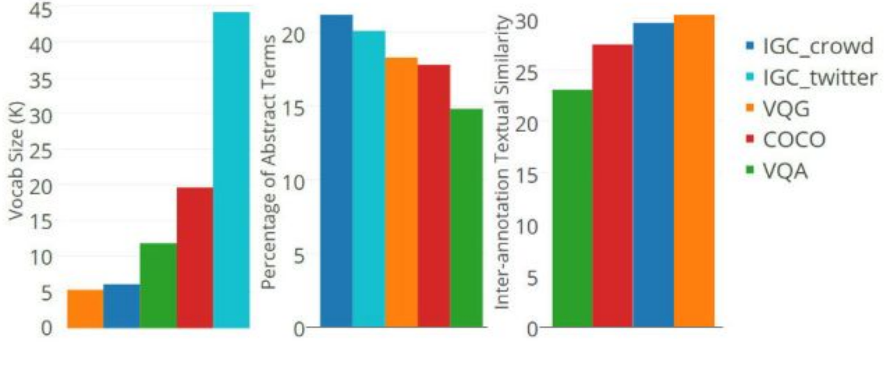


Figure 10: Comparison of V&L datasets.