



Beyond VQA: Generating Multi-word Answers and Rationales to Visual Questions

Radhika Dua*, Sai Srinivas Kancheti*, Vineeth N Balasubramanian

{radhika, cs21resch01004, vineethnb}@iith.ac.in

Indian Institute of Technology Hyderabad, India



Motivation

- Visual Question Answering is a multi-modal task that aims to measure high-level visual understanding.
- Contemporary VQA models are restrictive in the sense that answers are obtained via classification over a limited vocabulary (in the case of open-ended VQA), or via classification over a set of multiple-choice-type answers.
- There can be many correct multi-word answers to a question which makes this classification setting restrictive.

Input Image

Answer choices from VCR dataset

Question: Why are **person3** and **person4** sitting at the table?

a) **person3** and **person4** are having tea

b) they are eating dinner

c) **person3** and **person4** are waiting for breakfast

d) **person3** and **person4** are celebrating a birthday

Other possible multi-word answers:

- person3** and **person4** are in a party
- They are having drinks in a birthday party
- person3** and **person4** are attending a friend's birthday party
- They are talking to each other at a party

Task Description

- We introduce a new task: **ViQAR** (Visual Question Answering and Reasoning), wherein a model must generate the complete answer and a rationale that seeks to justify the generated answer.

Question: Why is **person1** covering his face ?

Generated answer: he is trying to avoid getting burned

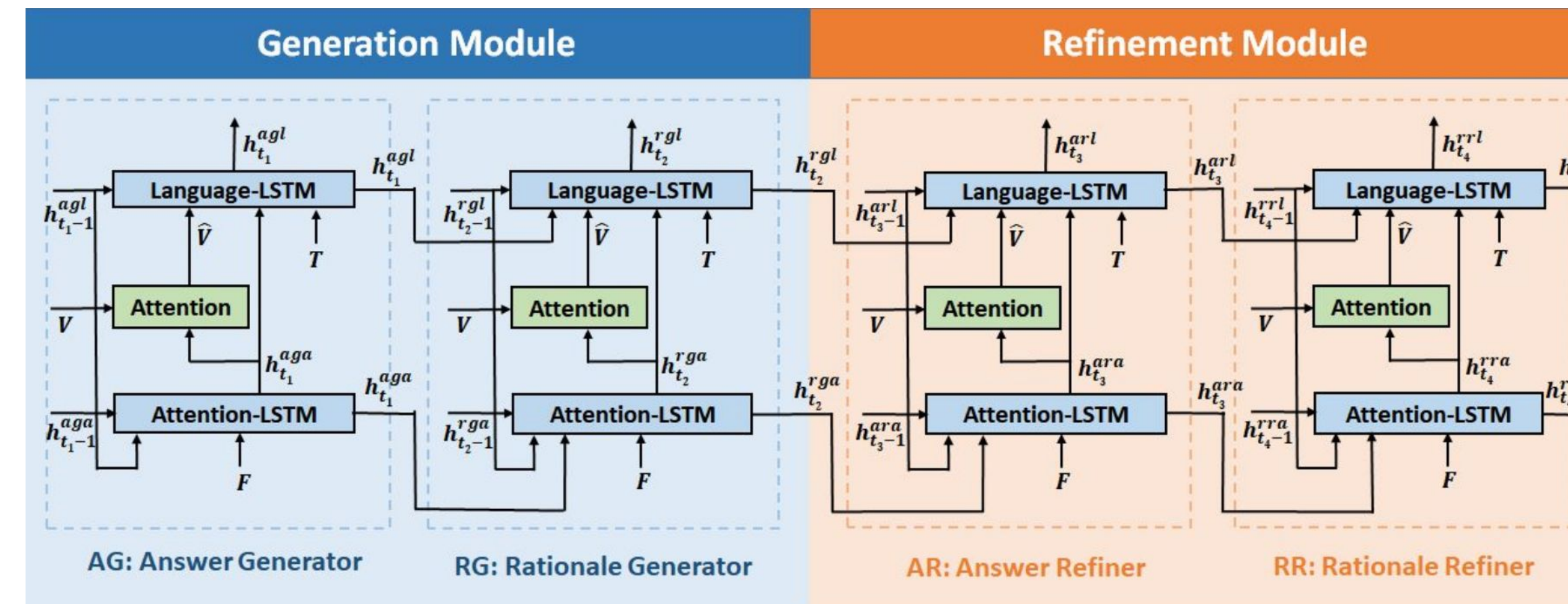
Why this answer?

Generated reason: there is a fire right in front of him

- The above example also illustrates the kind of visual question for which a **single-word answer is insufficient**.

Methodology

- Humans often use a rationale to answer a question, and sometimes vice versa. This suggests a close interplay between answer and rationale. Inspired by this interplay, we propose a model for answer and rationale generation.
- Our architecture consists of the *Generation Module* and the *Refinement Module*, both comprising of two sequential, stacked LSTMs.
- Simplicity of our approach suggests the tractability of our task.



Qualitative Results

Example output from our proposed model:

Question: What does **person2** do?

GT Answer: she is a student

GT Reason: she is wearing a school uniform

Generated Answer: **person2** is a student

Generated Reason: **person2** is wearing a school uniform

Challenging input for which our model fails:

Question: What is **person1** doing ?

GT Answer: They are cliff diving .

GT Reason: **person1** is wearing only shoes and trunks . he seems to be jumping down to the water from a high place .

Generated Answer: **person1** is performing a dance

Generated Reason: **person1** is standing on the stage with his arms raised

Quantitative Results

- We compare our proposed model and its variants against a basic two-stage LSTM model and a VQA model [1] as baselines [CS = cosine similarity].

Metrics	VQA-Baseline	Baseline	Q+I+C (Ours)	Q+I (Ours)	Q+C (Ours)
Univ Sent Encoder CS	0.419	0.410	0.455	0.454	0.440
Inferent CS	0.370	0.400	0.438	0.442	0.426
Embedding Avg CS	0.838	0.840	0.846	0.853	0.845
Vector Extrema CS	0.474	0.444	0.493	0.483	0.475
Greedy Matching Score	0.662	0.633	0.672	0.661	0.657
METEOR	0.107	0.095	0.116	0.104	0.103
Skipthought CS	0.430	0.359	0.436	0.387	0.385
RougeL	0.259	0.206	0.262	0.232	0.236
CIDEr	0.364	0.158	0.455	0.310	0.298
F-BERTScore	0.877	0.860	0.879	0.867	0.868

- We also perform a human Turing test on the generated answers and rationales, and find that our model generates grammatically correct and consistent answers and rationales.

Ablation Results

- We compare our proposed architecture with variations in number of refinement modules.

Image	Question	Generation Module	Generation - Refinement Module	Metrics	#Refine Modules		
					0	1	2
	Where are they at?	Answer: they are in a library Reason: there are shelves of books behind them	Answer: they are in a library Reason: there are shelves of books behind them	Univ Sent Encoder CS	0.453	0.455	0.430
	What are person1, person2, person3, person4, and person5 doing here?	Answer: they are studying a class Reason: they are all sitting in a circle and there is a teacher in front of them	Answer: they are all to attend a funeral Reason: they are all wearing black	Inferent CS	0.434	0.438	0.421
				Embedding Avg CS	0.850	0.846	0.840
				Vector Extrema CS	0.482	0.493	0.462
				Greedy Matching Score	0.659	0.672	0.639
				METEOR	0.101	0.116	0.090
				Skipthought CS	0.384	0.436	0.375
				RougeL	0.234	0.262	0.198
				CIDEr	0.314	0.455	0.197
				F-BertScore	0.868	0.879	0.861

[1] Peter Anderson et al. "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering". In: *CVPR*. 2018.