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

Radhika Dua[1], Sai Srinivas Kancheti[1] and Vineeth N Balasubramanian

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

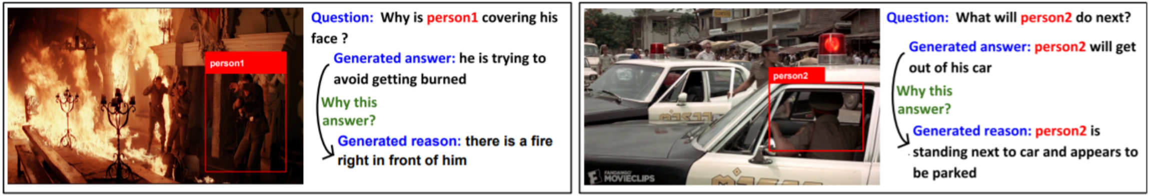
Indian Institute of Technology Hyderabad, India

4th Multimodal Learning and Applications Workshop in conjunction with CVPR 2021

[1] Equal contribution

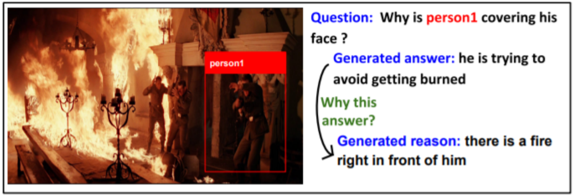
ViQAR: Task Overview

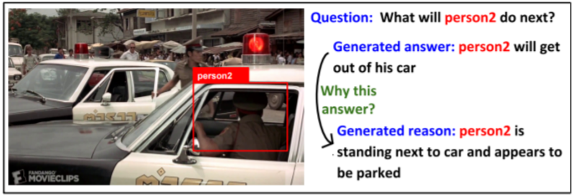
Given an image and a question about the image, we **generate a natural language answer and reason** that explains why the answer was generated.



ViQAR: Task Overview

Given an image and a question about the image, we **generate a natural language answer and reason** that explains why the answer was generated.

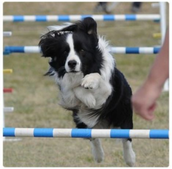




These examples also illustrate the kind of visual questions for which a **single-word answer is insufficient**.

ViQAR: Introduction

Vision-Language tasks:



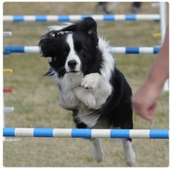
"black and white dog jumps over bar."

Image Captioning

A. Karpathy and Li Fei-Fei. “Deep Visual-Semantic Alignments for Generating Image Descriptions”. In: *IEEE*

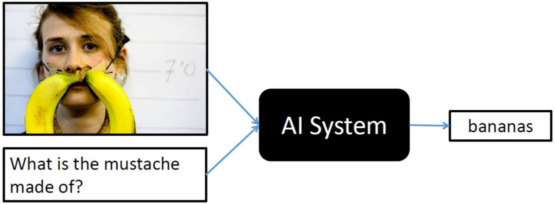
ViQAR: Introduction

Vision-Language tasks:



"black and white dog jumps over bar."

Image Captioning



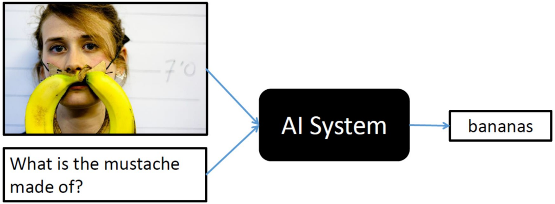
Visual Question Answering

Stanislaw Antol et al. "VQA: Visual Question Answering". In: *ICCV*. 2015.

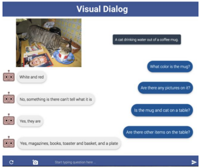
ViQAR: Introduction

Vision-Language tasks:

Image Captioning



Visual Question Answering

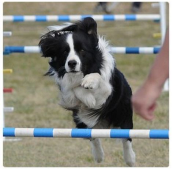


Visual Dialog

Abhishek Das et al. "Visual Dialog". In: *CVPR*. 2017.

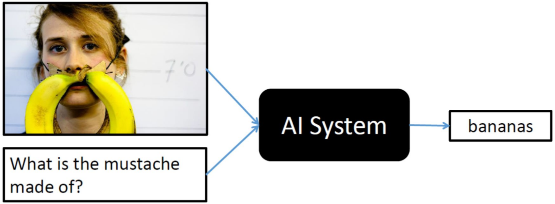
ViQAR: Introduction

Vision-Language tasks:

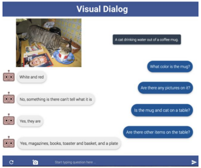


"black and white dog jumps over bar."

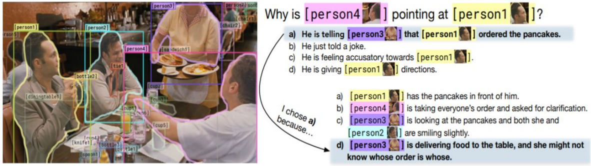
Image Captioning



Visual Question Answering



Visual Dialog

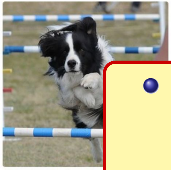


Visual Commonsense Reasoning

Rowan Zellers et al. "From Recognition to Cognition: Visual Commonsense Reasoning". In: *CVPR*. 2018.

ViQAR: Introduction

Vision-Language tasks:

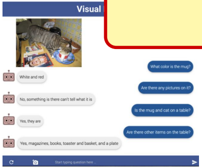


"black and white do
bar."

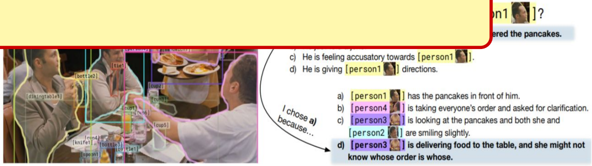
Image Capti

bananas

- Almost all of these tasks focus on **classifying an outcome** among a vocabulary of options.



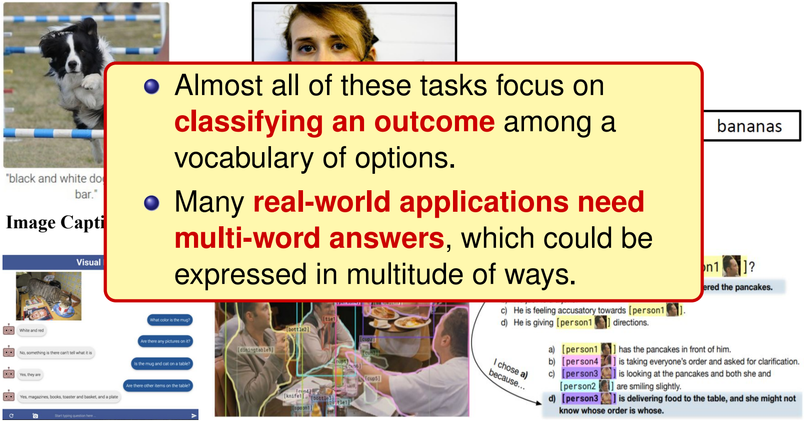
Visual Dialog



Visual Commonsense Reasoning

ViQAR: Introduction

Vision-Language tasks:



Visual Dialog

Visual Commonsense Reasoning

- Almost all of these tasks focus on **classifying an outcome** among a vocabulary of options.
- Many **real-world applications need multi-word answers**, which could be expressed in multitude of ways.

ViQAR: Proposed Architecture

- **Humans often use a rationale to answer** a question, and sometimes vice versa.

ViQAR: Proposed Architecture

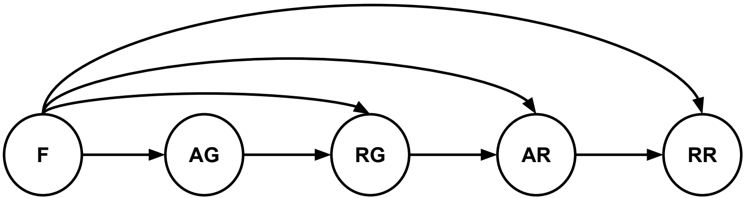
- **Humans often use a rationale to answer** a question, and sometimes vice versa.
- This suggests a **close interplay between answer and rationale**.

ViQAR: Proposed Architecture

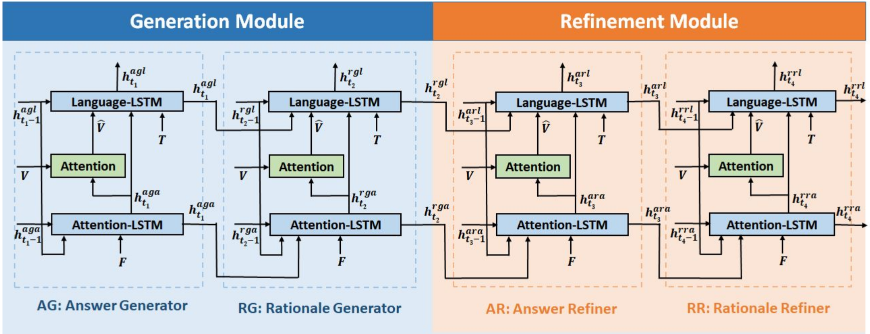
- **Humans often use a rationale to answer** a question, and sometimes vice versa.
- This suggests a **close interplay between answer and rationale**.
- Inspired by this interplay, we propose a model for answer and rationale generation.

ViQAR: Proposed Architecture

- **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.

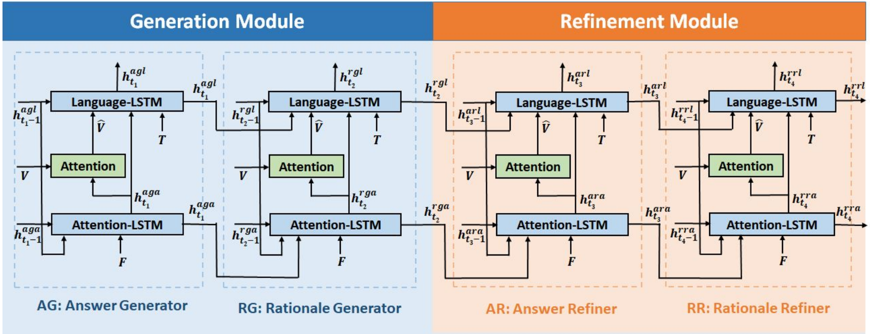


ViQAR: Proposed Architecture



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

ViQAR: Proposed Architecture



$$\text{Overall loss} = -\left(\sum_{t=1}^{l_a} \log p_t^{\theta_1} + \sum_{t=1}^{l_r} \log p_t^{\theta_2} + \sum_{t=1}^{l_a} \log p_t^{\theta_3} + \sum_{t=1}^{l_r} \log p_t^{\theta_4}\right)$$

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

ViQAR: 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

ViQAR: 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

A 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

ViQAR: Quantitative Results

We compare our proposed model and its variants against a basic two-stage LSTM model and a VQA model[2] as baselines.[CS = cosine similarity]

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

ViQAR: Quantitative Results

We compare our proposed model and its variants against a basic two-stage LSTM model and a VQA model[2] 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
Infersent 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

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

ViQAR: Ablation

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

ViQAR: Ablation

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

ViQAR: Ablation

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

Image		
Question	Where are they at?	What are person1, person2, person3, person4, and person5 doing here?

ViQAR: Ablation

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

Image		
Question	Where are they at?	What are person1, person2, person3, person4, and person5 doing here?
Generation Module	Answer: they are in a library Reason: there are shelves of books behind them	Answer: they are studying a class Reason: they are all sitting in a circle and there is a teacher in front of them

ViQAR: Ablation

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

Image		
Question	Where are they at?	What are person1, person2, person3, person4, and person5 doing here?
Generation Module	Answer: they are in a library Reason: there are shelves of books behind them	Answer: they are studying a class Reason: they are all sitting in a circle and there is a teacher in front of them
Generation - Refinement Module	Answer: they are in a liquor store Reason: there are shelves of liquor bottles on the shelves	Answer: they are all to attend a funeral Reason: they are all wearing black

ViQAR: Ablation

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

Image		
Question	Where are they at?	What are person1, person2, person3, person4, and person5 doing here?
Generation Module	Answer: they are in a library Reason: there are shelves of books behind them	Answer: they are studying a class Reason: they are all sitting in a circle and there is a teacher in front of them
Generation - Refinement Module	Answer: they are in a liquor store Reason: there are shelves of liquor bottles on the shelves	Answer: they are all to attend a funeral Reason: they are all wearing black

Metrics	#Refine Modules		
	0	1	2
Univ Sent Encoder CS	0.453	**0.455**	0.430
Infersent 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

ViQAR: Results

- We also perform **a human Turing test** on the generated answers and rationales.

ViQAR: Results

- We also perform **a human Turing test** on the generated answers and rationales.
- **30 human evaluators** were presented each with **50 randomly sampled image-question pairs**, out of which sixteen had ground truth answers and rationales.

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

ViQAR: Results

- We also perform **a human Turing test** on the generated answers and rationales.
- **30 human evaluators** were presented each with **50 randomly sampled image-question pairs**, out of which sixteen had ground truth answers and rationales.
- Evaluators had to give a **rating of 1 to 5**, with 1 being very poor and 5 being very good.

ViQAR: Results

- We also perform **a human Turing test** on the generated answers and rationales.
- **30 human evaluators** were presented each with **50 randomly sampled image-question pairs**, out of which sixteen had ground truth answers and rationales.
- Evaluators had to give a **rating of 1 to 5**, with 1 being very poor and 5 being very good.

Criteria	Generated	Ground-truth
How well-formed and grammatically correct is the answer?	4.15±1.05	4.40±0.87
How well-formed and grammatically correct is the rationale?	3.53±1.26	4.26±0.92
How relevant is the answer to the image-question pair?	3.60±1.32	4.08±1.03
How well does the rationale explain the answer with respect to the image-question pair?	3.04±1.36	4.05±1.10
Irrespective of the image-question pair, how well does the rationale explain the answer ?	3.46±1.35	4.13±1.09

Conclusion

- We propose ViQAR, a novel task for generating a multi-word answer and a rationale given an image and a question.

[3] Rowan Zellers et al. "From Recognition to Cognition: Visual Commonsense Reasoning". In: *CVPR*. 2018.

Conclusion

- We propose ViQAR, a novel task for generating a multi-word answer and a rationale given an image and a question. Our work aims to go beyond classical VQA by moving to a completely generative paradigm.

[3] Rowan Zellers et al. "From Recognition to Cognition: Visual Commonsense Reasoning". In: *CVPR*. 2018.

Conclusion

- We propose ViQAR, a novel task for generating a multi-word answer and a rationale given an image and a question. Our work aims to go beyond classical VQA by moving to a completely generative paradigm.
- We also present an end-to-end generation-refinement architecture which is based on the observation that answers and rationales are dependent on one another.

[3] Rowan Zellers et al. "From Recognition to Cognition: Visual Commonsense Reasoning". In: *CVPR*. 2018.

Conclusion

- We propose ViQAR, a novel task for generating a multi-word answer and a rationale given an image and a question. Our work aims to go beyond classical VQA by moving to a completely generative paradigm.
- We also present an end-to-end generation-refinement architecture which is based on the observation that answers and rationales are dependent on one another.
- We showed the promise of our model on the VCR dataset[3] both qualitatively and quantitatively, and our human Turing test showed results comparable to the ground truth.

[3] Rowan Zellers et al. "From Recognition to Cognition: Visual Commonsense Reasoning". In: *CVPR*. 2018.

Conclusion

- We propose ViQAR, a novel task for generating a multi-word answer and a rationale given an image and a question. Our work aims to go beyond classical VQA by moving to a completely generative paradigm.
- We also present an end-to-end generation-refinement architecture which is based on the observation that answers and rationales are dependent on one another.
- We showed the promise of our model on the VCR dataset[3] both qualitatively and quantitatively, and our human Turing test showed results comparable to the ground truth.
- We also showed that this model can be transferred to tasks without ground truth rationale.

[3] Rowan Zellers et al. "From Recognition to Cognition: Visual Commonsense Reasoning". In: *CVPR*. 2018.

Conclusion

- We propose ViQAR, a novel task for generating a multi-word answer and a rationale given an image and a question. Our work aims to go beyond classical VQA by moving to a completely generative paradigm.
- We also present an end-to-end generation-refinement architecture which is based on the observation that answers and rationales are dependent on one another.
- We showed the promise of our model on the VCR dataset[3] both qualitatively and quantitatively, and our human Turing test showed results comparable to the ground truth.
- We also showed that this model can be transferred to tasks without ground truth rationale.
- We hope that our work will open up a broader discussion around generative answers in VQA and other deep neural network models in general.

[3] Rowan Zellers et al. "From Recognition to Cognition: Visual Commonsense Reasoning". In: *CVPR*. 2018.