CITY
UNIVERSITY OF LONDON
EST 1894

The 17th Conference of the European Chapter of the Association for Computational Linguistics

Towards a Unified Model for Generating Answers and Explanations in Visual Question Answering

Chenxi Whitehouse, Tillman Weyde and Pranava Madhyastha (chenxi.whitehouse@city.ac.uk)

City, University of London

May 2023

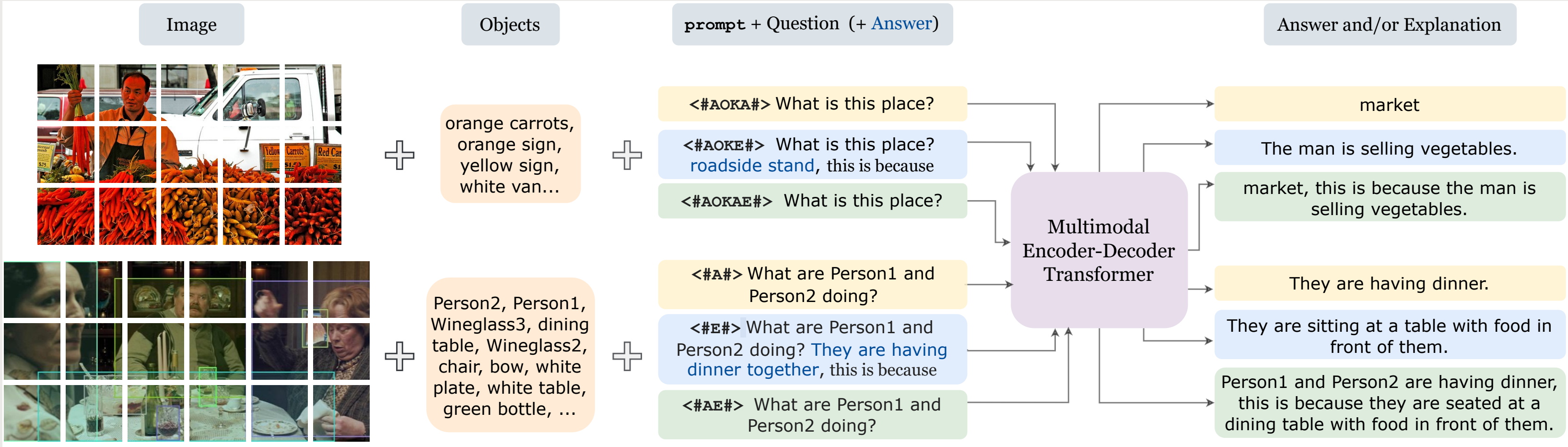
Motivation and Contribution

- Background:
  - Providing explanations for Visual Question Answering (VQA) tasks is desirable
  - Current explanation models for VQA generally trained separately from the QA model, resulting in less grounded answer and explanation

- Our proposal:
  - UMAE: A **u**nified **m**odel for **a**nswer and **e**xplanation generation
  - Multitask learning with single artificial prompt tokens to distinguish tasks while joint training
  - Use perplexity as criteria to map open-ended generations to multiple-choice options
  - SOTA explanation generation scores and promising out-of-domain performance on VQA

UMAE Illustration

- Train a multimodal encoder-decoder model on mix of VQA tasks for jointly optimizing answer & explanation
- Distinguish the training instances and target output with artificial prompt tokens (e. g. `<#A#>`, `<#E#>`).
- Top and bottom examples are from A-OKVQA (*Schwenk et al., 2022*) and VCR (*Zellers et al., 2019*), respectively



Artificial Prompt Tokens

- Add a single artificial prompt token at the beginning of the textual input to
  - Distinguish different datasets and tasks
    - <#A#> for generating answer, <#AE#> for explanations, <#AE#> for both answer and explanation
  - Learn the shared semantics among different tasks
- These tokens are abstract, simple yet effective
- Different from natural language prompt commonly used in seq2seq models such as T5

Artificial prompts

<#A#> What are Person1 and Person2 doing?

<#E#> What are Person1 and Person2 doing? They are having dinner together, this is because

<#AE#> What are Person1 and Person2 doing?

Natural Language prompts

[Provide answer and explanation] What are Person1 and Person2 doing?

Perplexity as Multiple-Choice Metric

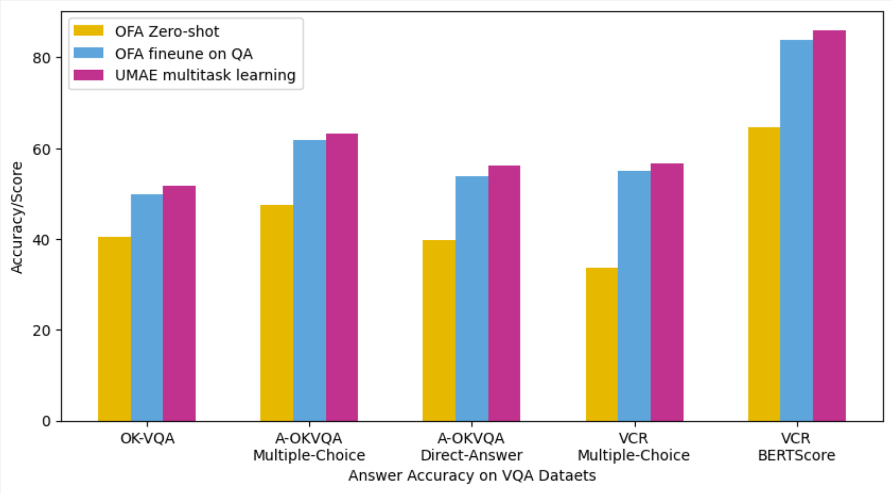
- Map open-ended generated text to multiple-choice options
  - Limitation of existing methods using semantic embedding similarities such as Glove
  - We instead feed the same visual and textual input to the model and calculate the perplexity of each answer being generation
  - Choose the lowest perplexity option as the final answer
  - Results in better performance than mapping with generation metrics (BLEU, BERTScore)

Model and Datasets

- We built on OFA, a multimodal encoder-decoder model *(Wang et al., 2022)*
  - Additionally extract bottom-up top-down features and attributes and feed to OFA
  - We do not use candidate answer set as OFA
- Datasets:
  - Train on three knowledge-intensive visual question answering tasks:
    - OKVQA (*Marino, et al., 2019*, question and answer)
    - A-OKVQA (answer and explanation)
    - VCR (answer and explanation)
  - Out-of-domain evaluation on VQA-X *(Park, et al., 2018)*

Experimental Results

- Answer Accuracy
  - UMAE achieves better results than trained models separately
  - Refer to the paper for more detailed scores



Experimental Results

- Explanation Performance
  - OFA zero-shot not able to follow natural language instruction to generate explanations
  - UMAE achieves SOTA explanation generation on A-OKVAQA, VCR and promising out-of-domain results on VQA-X

Dataset	Model	e-ViL Scores			N-gram Scores					Learnt Score
		S_O	S_T	S_E	BLEU4	ROUGE-L	METEOR	CIDEr	SPICE	BERTSCORE
A-OKVQA	OFA*	4.44	56.19	7.90	0.30	4.45	3.26	4.82	4.62	68.64
	$OFA_{Q\text{->}A}$+$OFA_{QA\text{->}E}$	35.82	74.32	48.29	22.18	48.51	23.56	86.76	22.46	85.96
	$UMAE_{A\text{-}OKVQA}$	37.10	73.97	50.15	**27.61**	52.23	24.06	**104.39**	22.88	87.86
	$UMAE_{ALL}$	**37.91**	**74.59**	**50.82**	27.35	**52.56**	**24.83**	101.09	**23.33**	**88.21**
VCR	e-UG	19.30	**69.80**	27.60	4.30	22.50	11.80	32.70	12.60	79.00
	$UMAE_{VCR}$	22.57	**56.68**	39.82	12.25	28.87	16.67	**48.14**	27.36	81.77
	$UMAE_{ALL}$	**22.82**	56.66	**40.27**	**13.44**	**29.53**	**17.54**	47.33	**26.45**	**81.91**
VQA-X	e-UG	36.50	80.50	45.40	23.20	45.70	22.10	74.10	20.10	87.00
	$UMAE_{ALL}$	31.58	77.65	40.67	14.63	35.12	20.29	50.35	19.13	85.40

Table 2: Explanation Scores. OFA* is the pretrained OFA, showing the transferability of OFA for generating explanations with natural language instructions. Results with e-UG are from Kayser et al. (2021). We show the best results of A-OKVQA and VCR in bold. The last row in blue shade shows *out-of-domain* performance.

Conclusion

- Jointly optimising answer and explanation improves quality in both in VQA
- Artificial prompt tokens is a simple and effect addition to the training data to boost multitask learning
- Perplexity as multiple-choice options metric outperform other metrics based on evaluating similarities
- We also discuss dataset quality limitation in the paper


Thank you!