JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

西北工業大學
NORTHWESTERN POLYTECHNICAL UNIVERSITY

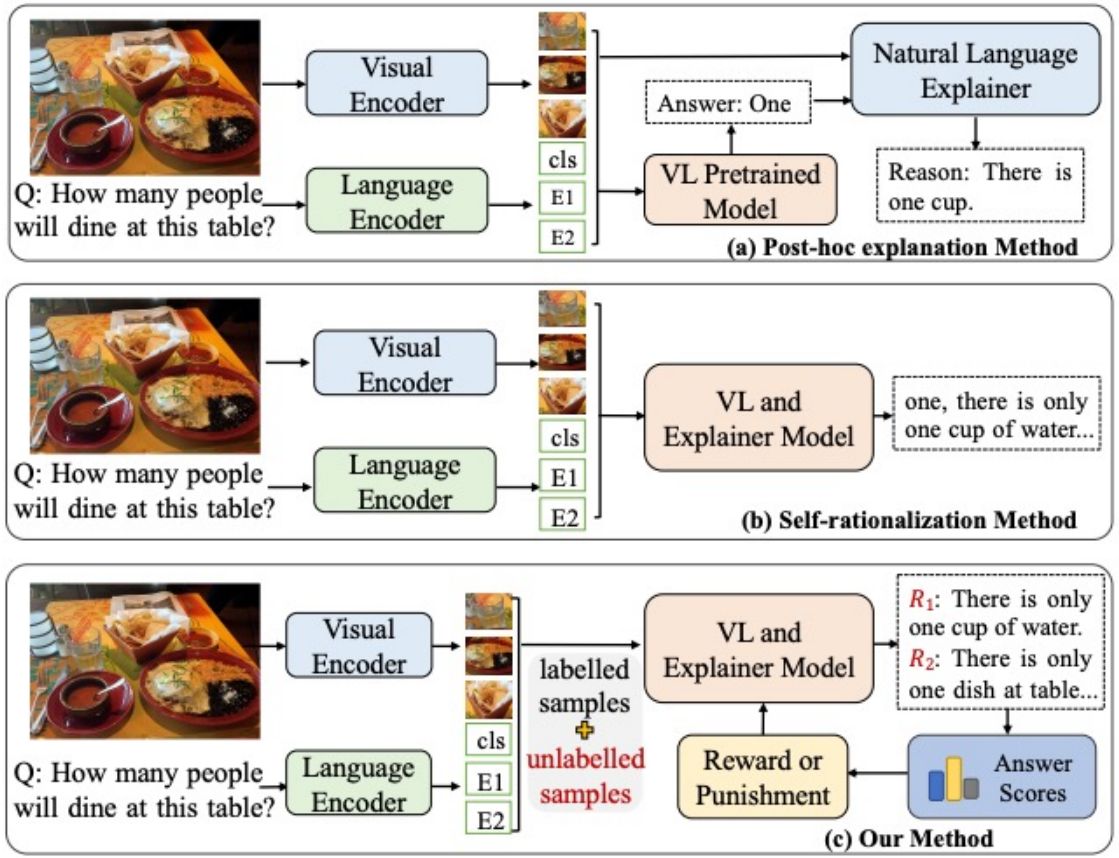
THE UNIVERSITY of ADELAIDE

S^{3}C: Semi-Supervised VQA Natural Language Explanation via Self-Critical Learning

Wei Suo, Mengyang Sun, Weisong Liu, Yiqi Gao,
Peng Wang*, Yanning Zhang and Qi Wu

1. Motivation and Contribution

1.1 Motivation



Motivation

1. Post-hoc explanation methods would lead to unfaithful responses, since the decision-making model and interpretation part are two separate modules.

2. Self-rationalization frameworks suffer from the problem of logical inconsistency.

3. These strategies all require an amount of human annotated explanations, which are expensive and time consuming to collect.

1.2 Contribution

The main contributions of our paper are as follows:

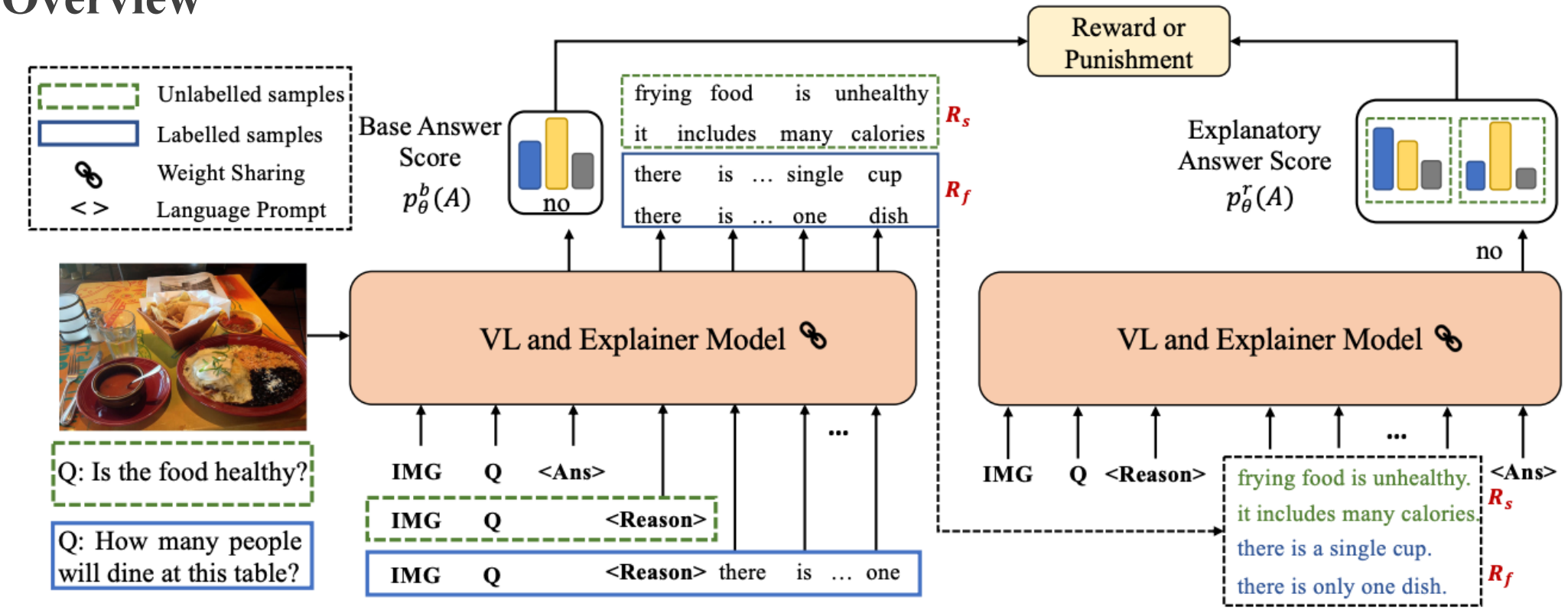
1. We propose a new self-critical VQA-NLE method that can model the logical relationships between answer explanation pairs and evaluate the generated rationales by answering rewards. This strategy effectively improves the logical consistency and the reliability of the interpretations.

2. We develop an advanced semi-supervised learning framework for VQA-NLE, which utilizes amounts of samples without human-annotated explanations to boost the self-interpretability of the model further. To the best of our knowledge, we are the first to explore semi-supervised learning on the VQA Natural Language Explanation.

3. The proposed S^3C achieves new state-of-the-art performance on VQA-X and A-OKVQA benchmarks. Meanwhile, automatic measures and human evaluations all show the effectiveness of our method.

2. Method

2.1 Overview



- we first use Answer-Explanation Prompt to obtain the base answer scores and candidate explanations with a pre-trained VL model.
- Then these reasons are reorganized and fed back into the model to capture the explanatory answer score.
- Further, our Self-Critical Reinforcement module evaluates the generated explanations and returns the rewards to improve the self-interpretability of the model.

3. Experiment Result

西北工業大學 NORTHWESTERN POLYTECHNICAL UNIVERSITY

3.1 Comparison with the SOTA methods on the two different datasets

Table 1. Comparison with the state-of-the-art methods on the **VQA-X**. Note that these results are **unfiltered** scores. S^3C^* denotes the model without unlabelled samples.

	VQA-X						
Approach	B4	M	R	S	C	Acc	Human
CAPS [41]	5.9	12.6	26.3	11.9	35.2	68.6	-
PJ-X [41]	19.5	18.2	43.4	15.1	71.3	76.4	65.4
FME [58]	24.4	19.5	47.7	17.9	88.8	75.5	-
NLX-GPT [48]	25.6	21.5	48.7	20.2	97.2	83.1	70.2
S^3C^*(ours)	26.5	22.0	49.0	20.9	100.5	83.7	73.9
S^3C (ours)	**27.8**	**22.8**	**50.7**	**21.5**	**104.4**	**85.6**	**77.4**

Table 2. Comparison with the state-of-the-art methods on the **A-OKVQA**. Note that these results are **unfiltered** scores. S^3C^* denotes the model without unlabelled samples.

	AOKVQA							
Approach	B4	M	R	S	C	Acc val	Acc test	Human
ViLBERT [29]	-	-	-	-	-	30.6	25.9	-
LXMERT [52]	-	-	-	-	-	30.7	25.9	-
KRISP [35]	-	-	-	-	-	33.7	27.1	-
Clipcap [49]	-	-	-	-	-	30.8	25.9	-
e-UG [20]	15.1	18.1	42.4	14.9	51.5	30.5	25.6	44.1
NLX-GPT [48]	20.1	17.0	46.3	15.8	65.4	32.7	28.7	46.9
S^3C^*(ours)	21.8	17.9	47.3	17.3	70.6	33.0	29.6	49.4
S^3C (ours)	**22.5**	**18.5**	**48.4**	**18.1**	**74.4**	**34.2**	**33.5**	**54.7**

- Our S3C outperforms both the post-hoc explanation methods [41, 58] and the self-rationalization method [48]. (The B4, M, R, S, C, Acc and Human are short for BLEU-4, METEOR, ROUGE-L, SPICE, CIDEr,)
- With semi-supervised paradigm, the results are further improved by 7.2 points on CIDEr indicator.

3.2 Ablation study

Table 4. **Main shortcomings.** The main shortcomings of unqualified explanations on the VQA-X dataset. For each sample, human evaluators can select multiple shortcomings.

Model	Irrelevant explanations	Insufficient explanations	Meaningless explanations
RVT [34]	25.7%	33.5%	11.4%
PJ-X [41]	21.1%	28.4%	9.2%
e-UG [20]	22.8%	25.4%	8.7%
NLX-GPT [48]	20.3%	22.2%	9.1%
S^3C (ours)	**17.3%**	**18.9%**	**8.2%**

Table 5. **Cross-dataset testing.** We alternately use the VQA-X and A-OKVQA as source dataset and target dataset to test the generalization of our framework.

	VQA-X→A-OKVQA					
Approach	B4	M	R	S	C	Acc
NLX-GPT [48]	10.7	12.7	34.2	10.7	35.4	10.4
S^3C (ours)	**12.0**	**13.3**	**34.3**	**12.5**	**45.3**	**18.8**
	A-OKVQA→VQA-X					
Approach	B4	M	R	S	C	Acc
NLX-GPT [48]	9.1	13.6	32.8	9.1	33.2	42.4
S^3C (ours)	**10.9**	**15.0**	**34.1**	**10.4**	**38.6**	**43.8**

Table 6. **Ablation study.** We ablate key components to demonstrate the effectiveness of our method. SCR and Semi are Self-Critical Reinforcement module and Semi-supervised learning paradigm respectively.

	question	image	answer	explanation	SCR	Semi	B4	M	R	S	C	Acc
1	✓	✓	✓	–	–	–	–	–	–	–	–	80.1
2	✓	✓	–	✓	–	–	24.4	20.7	47.3	19.5	90.4	–
3	✓	✓	✓	✓	–	–	27.5	22.9	50.4	21.9	109.1	82.2
4	✓	✓	✓	✓	✓	–	29.1	23.4	51.9	22.7	112.1	83.7
5	✓	✓	✓	✓	✓	✓	**30.7**	**23.9**	**52.1**	**23.0**	**116.7**	**85.6**

4.1 Qualitative analyses

(a) Is this man dressed formal?
NLX: [no] <he is wearing a tie>
Ours: [yes] <he is wearing a dress shirt and tie>
GT: [yes] <he's wearing a crisp dress shirt and a tie>

(b) Is he in a swimming pool?
NLX: [yes] <he is on a surfboard>
Ours: [no] <he is surfing in the ocean>
GT: [no] <the man is riding an ocean wave>

(c) What is the elephant doing?
NLX: [drinking] <it is in the water>
Ours: [drinking] <the trunk is in the water>
GT: [drinking] <his trunk is in the water>

(d) Are the people having a party?
NLX: [no] <there are no people present>
Ours: [yes] <they are laughing and having a good time>
GT: [yes] <they are standing in a kitchen with drinks in hand and laughing>

(e) Does this animal appear to be looking at the camera?
NLX: [yes] <it is looking at something>
Ours: [no] <it is looking away from the camera>
GT: [no] <the animal is looking off to the side>

(f) Is this a bar?
NLX: [yes] <there are people sitting at tables>
Ours: [yes] <the men are sitting at a bar with drinks in front of them>
GT: [yes] <there are drinks on a long countertop with people eating and drinking>

- We show the results of the state-of-the-art NLX-GPT, our method and ground-truth (they are short for NLX, Ours and GT) on the VQA-X.
- Through comparison, our model achieves better logical consistency between answers and explanations.

Thanks