QACE: Asking Questions to Evaluate an Image Caption

Hwanhee Lee[1], Thomas Scialom[2,3], Seunghyun Yoon[4], Franck Dernoncourt[4], Kyomin Jung[1]

[1]Seoul National University, Seoul, Korea
[2]Sorbonne Université, Paris, France
[3]reciTAL, Paris, France
[4]Adobe Research, San Jose, CA, USA

Image Caption Evaluation

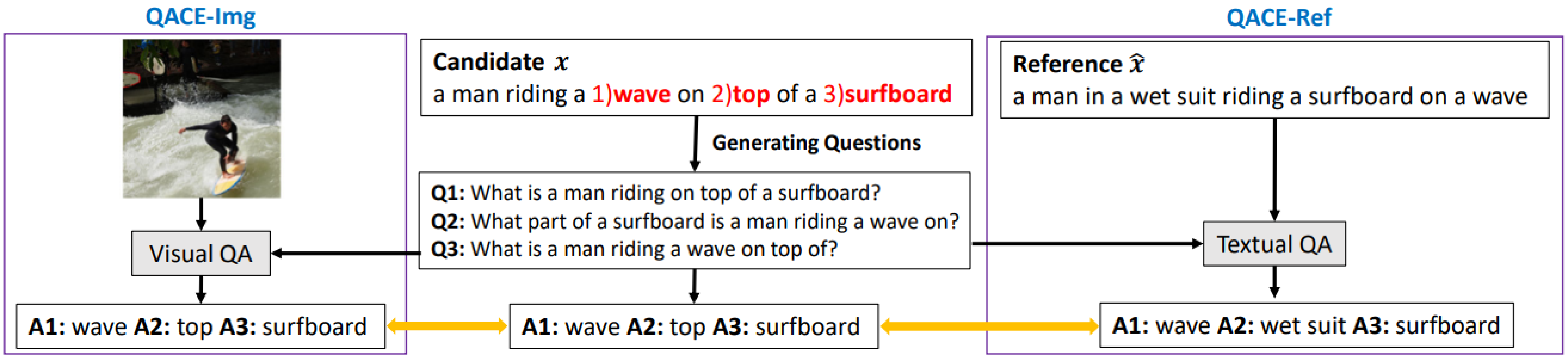
Reference: a passenger train pulled up to a covered platform with people standing on the platform.

Caption1: a blue subway train pulls into the subway station.

Caption2: a ***red*** train pulls into the platform.

- N-gram similarity metrics often fails to capture the semantic erros in the generated captions and require multiple references.

Overall Flow of QACE



① Extract possible answer spans (noun phrases) in a candidate caption.

② Generate answer-aware questions using answer spans and a candidate caption.

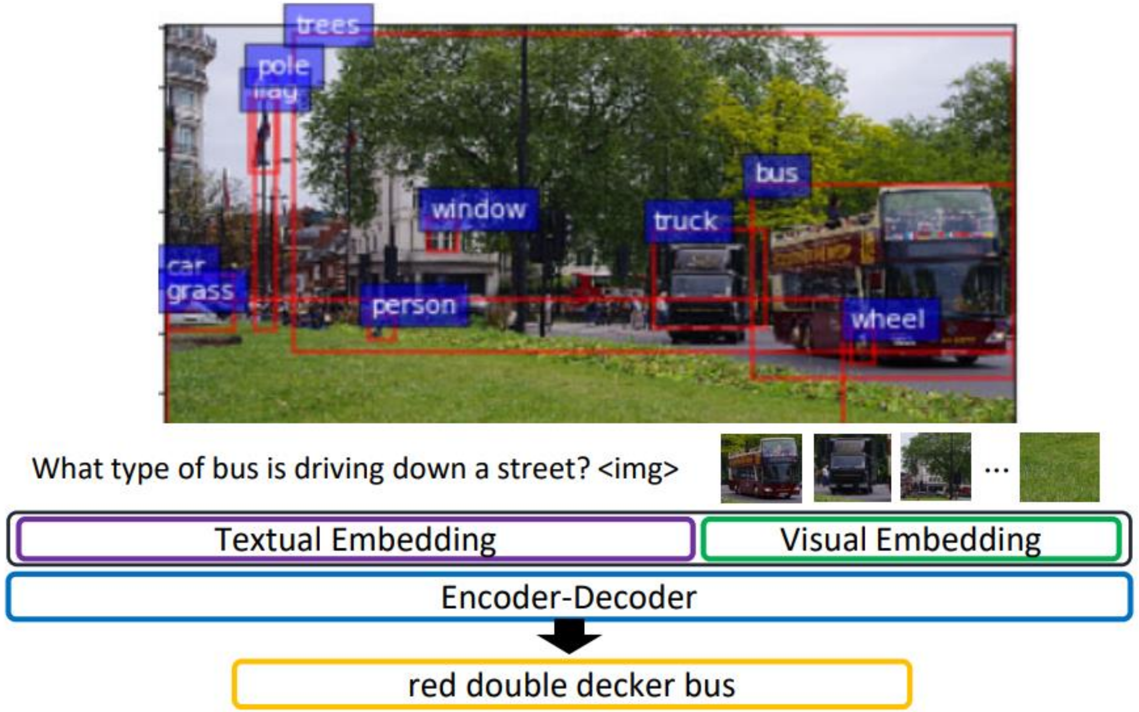
③ Generate answers using **“candidate caption”** and given context - **“image”** (VQA), **“reference”** (Textual QA)**.**

④ - QACE-Img: Compare the answers between an **“image”** and a **“candidate caption”.**

- QACE-Ref: Compare the answers between a **“reference caption”** and a **“candidate caption”.**

For comparing answers, we use *F1, BERTScore*, and *answerability.*

Abstractive VQA Model: Visual-T5



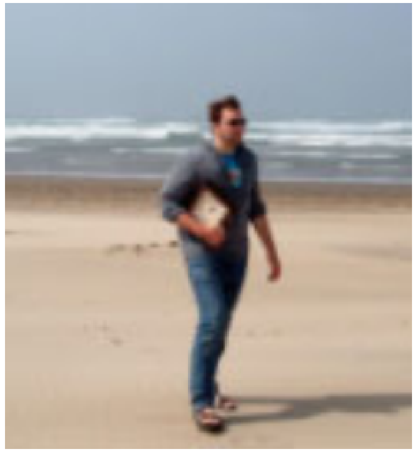
- Standard VQA models are framed as a classification among only a few thousand categories, and their usage is limited to comparing very few pre-defined answers.
- We propose an abstractive VQA system Visual-T5 as a new module for QACE-Img that can generate free-form abstractive answers given a textual question and an image.
- We conduct a human evaluation of Visual-T5 and show an accuracy of 69%.

Experimental Results

	Ref?	Pascal50s	Composite	Flickr8k
BLEU-4	✓	65.2	45.7	28.6
ROUGE-L	✓	67.7	47.7	30.0
METEOR	✓	**80.5**	46.6	40.3
CIDEr	✓	77.8	47.4	41.9
SPICE	✓	76.1	48.6	**45.7**
BERTScore	✓	72.0	45.6	30.5
QACE-Ref (ours)	✓	75.1	49.3	40.5
F1	✓	57.5	**55.1**	9.2
BERTScore	✓	76.4	46.0	30.9
Answerability	✓	71.6	47.3	39.0
-Perplexity	✗	46.8	1.7*	10.1
VIFIDEL	✗	69.0	13.1	**33.6**
QACE-Img (ours)	✗	**70.0**	**19.1**	29.1
F1	✗	62.0	12.5	27.3
BERTScore	✗	65.9	12.8	27.1
Answerability	✗	**74.5**	15.7	27.8

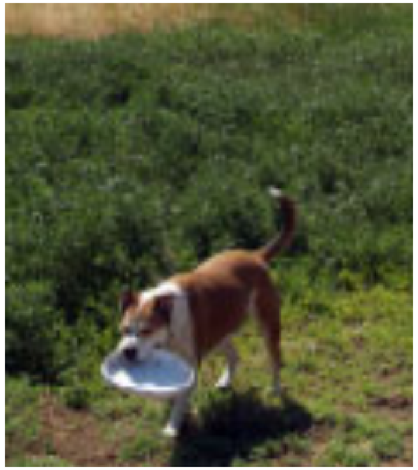
- We compute the correlation with human judgments for various metrics.
- Both QACE-Ref and QACE-Img show comparable or better performance than baseline metrics.
- Averaging the results of three answer similarity functions mostly show the best results.

Example



Candidate: a **man**^{A1} is standing on a **sunny beach**^{A2} **(Human: 1.0)**
Reference: a man walks down the beach near the ocean
Q1: What is standing on a sunny beach?
Q2: What is a man standing on?

Ref	**A1**:man **A2**:beach	$QAEC_{Ref}$: 0.88
Img	**A1**:man **A2**:sand	$QAEC_{Img}$: 0.79



Candidate: a **cow**^{A1} is standing in a **field**^{A2} of **grass**^{A3} **(Human: 0.2)**
Reference: a dog with a frisbee standing in the grass
Q1: What animal is standing in a field of grass?
Q2: What is a cow standing in?
Q3: What type of field is a cow standing in?

Ref	**A1**:dog **A2**:grass **A3**:grass	$QAEC_{Ref}$: 0.60
Img	**A1**:dog **A2**:unanswerable **A3**:grassy field	$QAEC_{Img}$: 0.47

Closing Remarks

- We propose a new captioning metric QACE, which generates questions on the evaluated caption and checks its content by asking the questions on either the reference caption (QACE-Ref) or the source image (QACE-Img).

- We propose *Visual-T5*, an abstractive VQA system that can generate free-from answers as a component of QACE-Img.

- Experimental results show that both QACE-Ref and QACE-Img show comparable or better performance than baseline metrics.

Code: https://github.com/hwanheelee1993/QACE