An Examination of the Robustness of Reference-Free Image Captioning Evaluation Metrics

Saba Ahmadi, Aishwarya Agrawal

Mila · Université de Montréal

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Reference-free metrics for image caption evaluation

1. These metrics utilize pretrained models to compute image-text similarity, which is then used as an evaluation score.
2. These metrics have been shown to perform better than n-gram based metrics.

But, are reference-free metrics robust enough?

Candidate Captions	CLIPScore	UMIC
The title of the book is topology.	0.472	0.347
The title of the book is muffin.	0.546	0.446

Reference-free Metrics

	CLIPScore [1]	UMIC [2]
Base Model	CLIP	UNITER
Fine-tuned	No	COCO + Negative samples

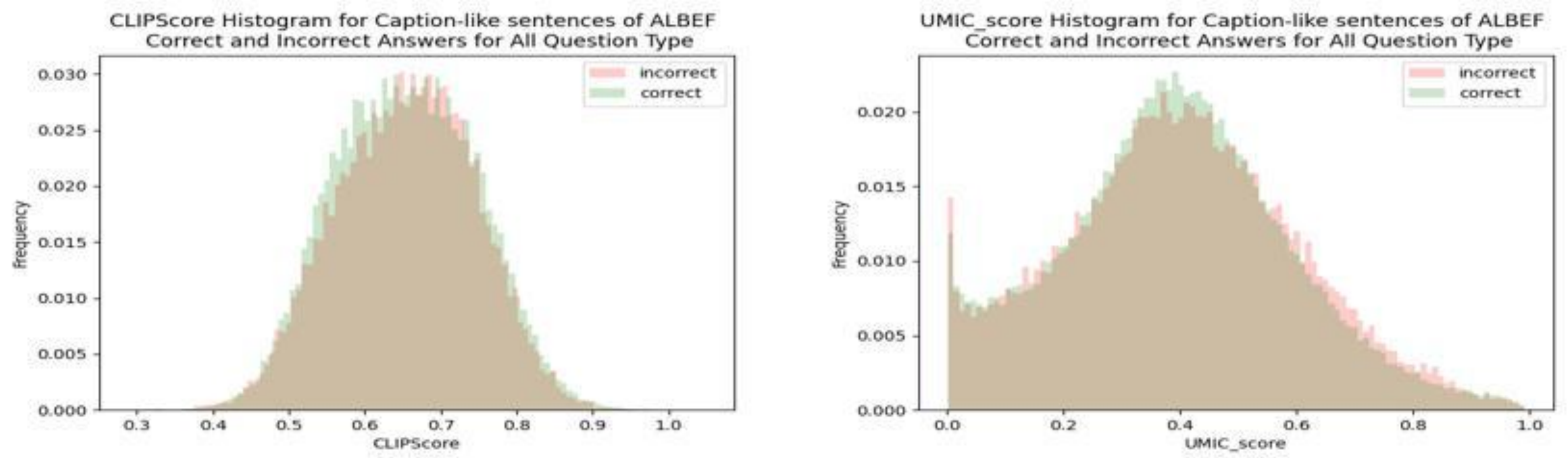
- Datasets used to conduct the examination

VQA datasets ([3], [4])
Question: What color is the tennis player's shirt? **Answer**: Red.
Converted caption: The color of tennis player's shirt is red.

COCO detection [5]**:** "There is a/an [object name]."

Preliminary Experiment

Incorrect and correct QA captions scores distribution



- The **significant overlap** in scores for correct and incorrect captions reveals these metrics' **limitations in accurately evaluating caption quality**.

Sensitivity to fine-grained errors

"Fine-grained error": Refers to error between a pair of correct and incorrect captions that have **high lexical overlap.**

- **Both** metrics **fail** to rank correct captions above incorrect captions when the **difference is fine-grained** in ~46% of times.

Are metrics differently sensitive to different kinds of fine-grained errors?

Answer Type	Example	CLIPScore	UMIC
Ground Truth	The color of the grass is green.	0.501±0.127	0.487±0.193
Plausible	The color of the grass is **white**.	0.474±0.124	0.242±0.181
Image Object	The color of the grass is **giraffe**.	0.526±0.119	0.354±0.154
Random	The color of the grass is **grill**.	0.458±0.124	0.275±0.160

- **Low** sensitivity to caption **implausibility.**
- **High** sensitivity to **visual grounding.**

Sensitivity to the number of objects

#Objects	Example	CLIPScore	UMIC
One Object	There is a person.	0.449±0.112	0.205±0.111
Two Objects	There is a person and a sports ball.	0.512±0.129	0.212±0.175
Three Objects	There is a person, a sports ball and a baseball bat.	0.561±0.129	0.195±0.175

- **CLIPScore** shows **high** sensitivity to the **number of image-relevant objects.**

Sensitivity to the size of objects

Object Size	Example	CLIPScore	UMIC
Small	There is a knife.	0.396±0.131	0.317±0.162
Big	There is a pizza.	0.434±0.134	0.232±0.138

- **Both** metrics **demonstrate sensitivity** to the **size** of image-relevant objects mentioned in the caption.

Sensitivity to negation

UMIC ranked the negated caption above the correct caption incorrectly in 44.24% of cases, while CLIPScore failed in 41.36% of cases.

- **Both** exhibited a **weak understanding** of **negation.**

Sensitivity to the length of caption

Caption type	Example	CLIPScore	UMIC
Short but Correct	There is a fire trucks bumper.	0.512±0.116	0.264 ±0.149
Long but Incorrect	The fire trucks bumper is made of plastic.	0.521±0.129	0.351±0.196

- **UMIC** exhibits **significant sensitivity** to caption **length**.

Sensitivity to sentence structure

CLIPScore fails to assign a higher score to the correct caption than the shuffled one in 34.32% of cases, whereas this occurs in only 9.18% of cases for UMIC.

- **UMIC** is **more responsive** to the **sentence structure.**

Conclusion

- We need to be cautious when deploying this metrics.
- We hope our findings will guide future research on developing more robust metrics.

References

[1] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021.
[2] Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Trung Bui, and Kyomin Jung. 2021.
[3] Kushal Kafle and Christopher Kanan. 2017.
[4] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016.
[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014.