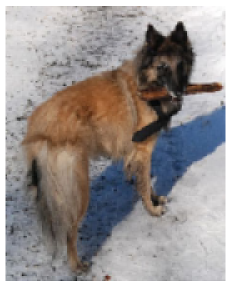
UMIC: An Unreferenced Metric for Image Captioning via Contrastive Learning

Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Trung Bui and Kyomin Jung
Seoul National University, Dept. of Electrical and Computer Engineering & Adobe Research

Association for Computational Linguistics
ACL-IJCNLP 2021

Problem



Ref 1: A dog standing in the snow with a stick in its mouth.
Ref 2: A little dog holding sticks in its mouth.
Candidate: A dog standing on the snow with a dog
CIDEr with Ref 1: 3.166
CIDEr with Ref 2: 0.281
Human Judgments : 1.875 out of 5

- The metric score for a given candidate caption varies significantly depending on the reference type due to the diverse nature of image captions.
- Reference-based metrics usually require multiple references, which are difficult to obtain, to get meaningful score.

Contributions

- We **introduce a new metric UMIC, an Unreferenced Metric for Image Captioning** which does not require reference captions to evaluate image captions based on Vision-and-Language BERT
- We observe critical problems of the previous benchmark dataset (i.e., human annotations) on image captioning metric, and **introduce a new collection of human annotations, *CapEval1k*, on the generated captions**
- We validate UMIC on four datasets, including our new dataset, and show that **UMIC has a higher correlation than most of the previous metrics that require multiple references.**

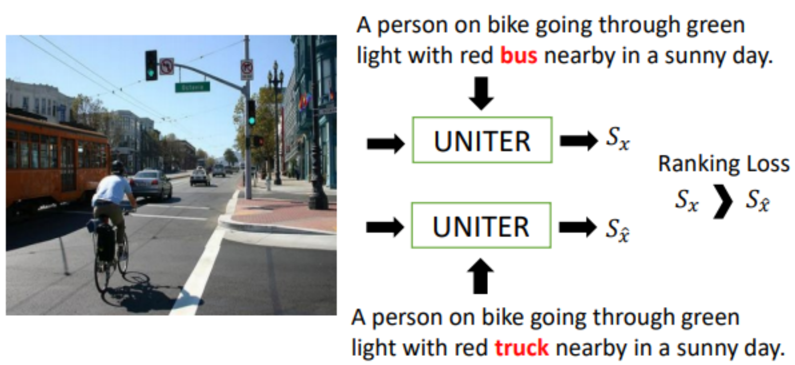
Generating Negative Captions

- We **prepare the negative captions that can represent most of the undesirable cases in captioning**, such as *relevant but have wrong keyword, irrelevant to the image, grammatically incorrect.*

1) **Substituting Keywords:** substitute 30% of the words(verb, adjective, noun) in the reference captions
2) **Random Captions:** sample captions from other images utilize the captions of the images similar to the given images
3) **Repetition & Removal :** repeat or remove some words in the reference captions with a probability of 30%
4) **Word Order Permutation:** changing the word order of the reference captions

Original: a woman hugging a girl who is holding a suitcase
Substitiution: a boy hugging a girl who is holding a suitcase
Random(Hard Negative): a very small cute child by a suitcase
Repetition & Removal: a woman hugging a girl is holding a suitcase suitcase

Training UMIC



- We **fine-tune UNITER via contrastive learning**, where the model is trained to compare and discriminate the ground-truth captions and diverse synthetic negative samples as follows.

1) $[CLS], i_1, \ldots, i_N, x_1, \ldots, x_T = \text{UNITER}(I, X)$.
2) $S(I, X) = sigmoid(W i_{[CLS]} + b)$
3) $Loss = max(0, M - (S(I, X) - S(I, \hat{X})))$

Comparison with Existing Metrics

Metric	Flickr8k	Composite	CapEval1k	PASCAL50s
BLEU-1	0.274	0.406	0.233	74.3
BLEU-4	0.286	0.439	0.238	73.4
ROUGE-L	0.300	0.417	0.220	74.9
METEOR	0.403	0.466	0.288	78.5
CIDEr	0.419	0.473	0.307	76.1
SPICE	0.457	0.486	0.279	73.6
BERTScore	0.396	0.456	0.273	79.5
BERT-TBR	0.467	0.439	0.257	**80.1**
VBTScore	**0.525**	**0.514**	**0.352**	79.6
VIFIDEL	0.336	0.191	0.143	70.0
UMIC	**0.468**	**0.561**	**0.328**	**85.1**
UMIC$_{-C}$	0.431	0.554	0.299	84.7

- We show that although UMIC does not utilize any reference captions, **UMIC outperforms most of the baseline metrics in all of the datasets that depend on multiple references.**

Example

References
- two giraffe standing next to each other in a field.
- two giraffes are climbing a hill with mountains in the background.

Candidate
- three giraffes standing in a field of grass

BLEU1: 0.324	**ROUGE-L**: 0.320	**METEOR**: 0.173	**CIDER**: 0.866
SPICE: 0.289	**UMIC**: 0.352	**UMIC**$_{/-c}$: 0.770	**Human**: 0.200

References
- a person breadking a bottle with a baseball bat
- a boy in yellow shirt swinging a baseball bat

Candidate
- a man swinging a baseball bat at a ball

BLEU1: 0.360	**ROUGE-L**: 0.354	**METEOR**: 0.176	**CIDER**: 1.205
SPICE: 0.192	**UMIC**: 0.094	**UMIC**$_{/-c}$: 0.062	**Human**: 0.450

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