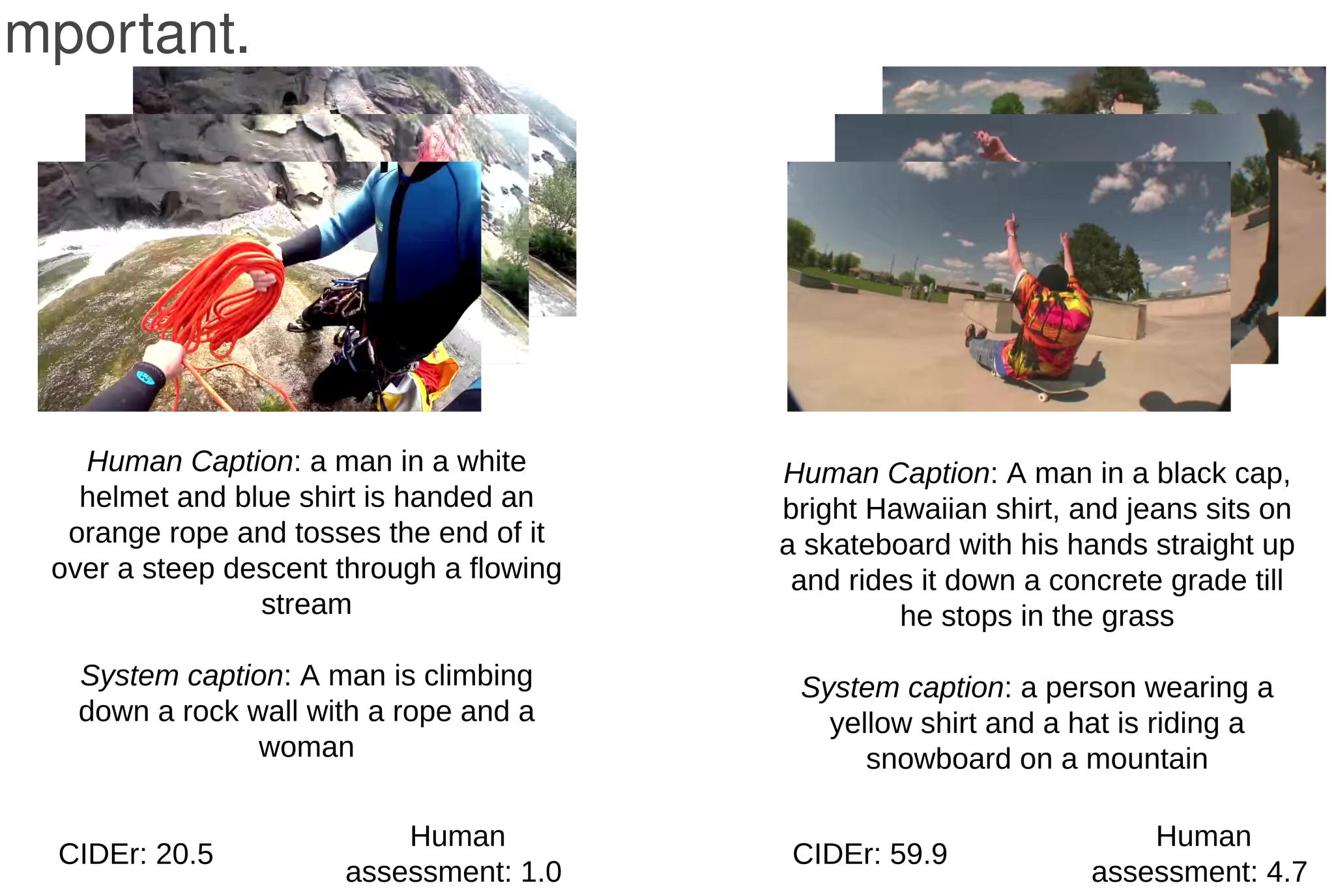
# BERTHA: Video Captioning Evaluation Via Transfer-Learned Human Assessment

Luis Lebron[1], Yvette Graham[2], Kevin McGuinness[1], Konstantinos Kouramas[3], Noel E. O'Connor[1]

Insight SFI Research Centre for Data Analytics, Dublin City University (DCU)[1]; School of Computer Science and Statistics, Trinity College Dublin[2]; Collins Aerospace[3]

## Introduction

Extracting textual descriptions is a difficult task by itself. In part, because is difficult to automatically evaluate if a caption is good automatically. Multiple factors need to be taken into account; for instance: the fluency of the caption, multiple actions happening in a single scene, and the human bias of what is considered important.



*Human Caption*: a man in a white helmet and blue shirt is handed an orange rope and tosses the end of it over a steep descent through a flowing stream

*System caption*: A man is climbing down a rock wall with a rope and a woman

CIDEr: 20.5 Human assessment: 1.0

*Human Caption*: A man in a black cap, bright Hawaiian shirt, and jeans sits on a skateboard with his hands straight up and rides it down a concrete grade till he stops in the grass

*System caption*: a person wearing a yellow shirt and a hat is riding a snowboard on a mountain

CIDEr: 59.9 Human assessment: 4.7

Figure 1: In these two pictures, we can see an example of captions where a popular automatic metric (CIDEr) fails to accurately evaluate the caption. In the first case, the system caption does not contain as much detail as the human caption, and it includes a woman who is nowhere in the scene. In the second case, the system caption seems like it does not describe the same scene; however, it contains some keywords in the human caption.

The contributions presented in this paper are as follows:

- A new video captioning metric trained in human judgments base on BERT [2].
- An evaluation of the performance of the metric and how it compares with other commonly used metrics in a challenging dataset, consisting of human judgments of the captions produced by the system participating in various years of the TRECVid video to text task [1]
- A study of the behaviour of metric under different scenarios and a test of the limits of the metric to understand its behaviour.

## Evaluation of Video Captioning via Transfer Learning

The goal of the model is to maximize the correlation, $\rho(\Gamma_\theta, \mathrm{A})$, between predicted scores ($\Gamma_\theta \in \mathbb{R}^{N' \times M' \times S}$) and the given human qualitative scores $\mathrm{A}$:

$$\underset{\theta}{\operatorname{argmax}} \ \rho(\Gamma_\theta, \mathrm{A}), \tag{1}$$

where $\rho$ is the Pearson product-moment correlation coefficient.

BERTHA is based on BERT and fine-tune in a dataset consisting on human judgments of video captions. A multilayer preceptor with a single output is attached to the $[CLS]$ symbol of BERT.

## Dataset

We use datasets from the past TRECVid [1] benchmark video to text task (VTT) from 2016, 2017, 2018, 2019 and 2020. Approximately 2,000 videos are available each year from Vine and later also from Flickr and V3C2.

The human judgements are divided into two sets: single annotator (SA) and multiple annotators (MA). The characteristics of these datasets are the following:

- 56,088 human judgments in the SA.
- 7,705 annotations in the MA set.
- SA has a mean of 15 tokens per human reference caption, and nine tokens of the system generate captions in terms of length.
- MA has a mean of 14 tokens per human reference and eight tokens per system-generated captions.

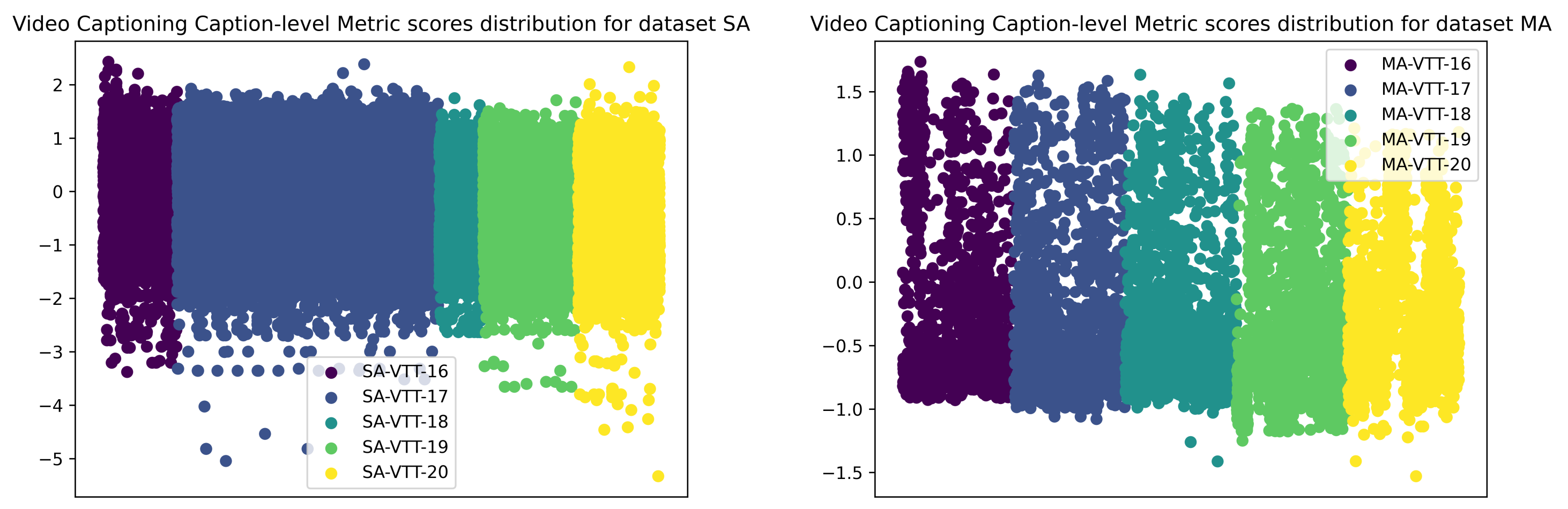


Figure 2: Score distribution of video captioning caption-level metric scores with human assessment SA(left) and MA(right) for TRECvid 2016 to 2020.

## Results

Two configuration use: BERTHA-SA and BERTHA-MA. Each model is trained in all the years of one of the datasets, e.g. BERTHA-SA is trained in the single annotator dataset. Each dataset is divided by each year to represent better the typical set-up of the TRECVid challenge.

| | SA-VTT-19 | SA-VTT-20 | MA-VTT-19 | MA-VTT-20 |
|---|---|---|---|---|
| BERTHA-SA | **0.929** | 0.963 | 0.859 | 0.888 |
| BERTHA-MA | 0.837 | 0.863 | 0.882 | 0.863 |
| BLEU-4 | 0.581 | 0.944 | 0.753 | 0.967 |
| CIDEr | 0.810 | **0.977** | **0.892** | 0.967 |
| METEOR | 0.887 | 0.958 | 0.880 | **0.988** |
| Rouge | 0.588 | 0.919 | 0.811 | 0.914 |
| SPICE | 0.498 | -0.110 | -0.240 | 0.146 |

Table 1: Pearson correlation of video captioning **system**-level metric scores with human assessment SA and MA for TRECVid 2019 and 2020 participating systems.

| | SA-VTT-19 | SA-VTT-20 | MA-VTT-19 | MA-VTT-20 |
|---|---|---|---|---|
| BERTHA-SA | **0.075** | 0.069 | 0.147 | 0.164 |
| BERTHA-MA | 0.072 | 0.032 | **0.225** | **0.247** |
| BLEU-4 | 0.030 | 0.010 | 0.049 | 0.112 |
| SentBLEU | 0.053 | 0.027 | -0.051 | 0.077 |
| CIDEr | 0.035 | 0.107 | 0.155 | 0.208 |
| METEOR | 0.064 | **0.115** | 0.222 | 0.235 |
| Rouge | 0.046 | 0.093 | 0.158 | 0.209 |
| SPICE | 0.032 | 0.003 | 0.018 | 0.001 |

Table 2: Pearson correlation of video captioning **caption**-level metric scores with human assessment SA and MA for TRECVid 2019 and 2020 participating systems.

In some cases, we need a single metric to define the performance. We train a linear model to find the best combination of metrics to predict human judgment to produce a new fusion metric. The best fit linear regression coefficients are:

| BERTHA | BLEU-4 | CIDEr | METEOR | ROUGE |
|---|---|---|---|---|
| 0.0525 | -0.1373 | 0.0315 | 0.2810 | -0.0779 |

## References

[1] George Awad et al. "TRECVID 2020: A comprehensive campaign for evaluating video retrieval tasks across multiple application domains". In: *Proceedings of TRECVID 2020* (2021).

[2] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

Acknowledgments

This research was supported by the Irish Research Council Enterprise Partnership Scheme together with United Technologies Research Center Ireland and the Insight SFI Research Centre for Data Analytics supported by Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2, co-funded by the European Regional Development Fund.

HOST INSTITUTIONS

PARTNER INSTITUTIONS

FUNDED BY:

Contact: luis.lebroncasas@insight-centre.org