

Evaluation metrics for video captioning: A survey

Andrei de Souza Inácio^{a,b,*}, Heitor Silvério Lopes^b

^a Federal Institute of Education, Science and Technology of Santa Catarina, Gaspar, 89111-009, SC, Brazil

^b Graduate Program in Electrical Engineering and Industrial Informatics, Federal University of Technology – Paraná, Curitiba, 80230-901, PR, Brazil

ARTICLE INFO

Keywords:

Automatic evaluation metrics
Learned metrics
Video captioning
Word embedding

ABSTRACT

Automatic evaluation metrics play an important role in assessing video captioning systems. Popular metrics used for assessing such approaches are based on word matching and may fail to evaluate the quality of automatically generated captions due to inherent natural language ambiguity. Moreover, they require many reference sentences for effective scoring. With the fast development of image and video captioning methodologies using deep learning in recent years, many metrics have been proposed for evaluating such approaches. In this study, we present a survey of automatic evaluation metrics for the video captioning task. Moreover, we highlight the challenges in evaluating video captioning and propose a taxonomy to organize the existing evaluation metrics. We also briefly describe and identify the advantages and shortcomings of those metrics and identify applications or contexts in which these metrics can be better used. To identify the advantages and limitations of the evaluation metrics, we quantitatively compare them using videos from different datasets employed for the video description task. Finally, we discuss the advantages and limitations of the metrics and propose some promising future research directions, such as semantic measurement, explainability, adaptability, extension to other languages, dataset limitations, and multimodal free-reference metrics.

Contents

1.	Introduction	2
2.	Video captioning vs image captioning	3
3.	Video captioning datasets	4
4.	Evaluation metrics	4
4.1.	Reference-based metrics	5
4.1.1.	BLEU (BiLingual Evaluation Understudy)	5
4.1.2.	METEOR (Metric for Evaluation of Translation with Explicit ORdering)	5
4.1.3.	CIDEr (Consensus-based Image Description Evaluation)	6
4.1.4.	ROUGE (Recall-Oriented Understudy for Gisting Evaluation)	6
4.1.5.	SPICE (Semantic Propositional Image Caption Evaluation)	6
4.1.6.	WMD (Word Mover's Distance)	6
4.1.7.	WEmbSim	7
4.1.8.	BERTScore (Bidirectional Encoder Representations from Transformers Score)	7
4.1.9.	SMURF (SeMantic and linguistic Understanding Fusion)	7
4.1.10.	VIFIDEL (VISual Fidelity for Image Description Evaluation)	8
4.1.11.	TIGER (The Text-to-Image Grounding based metric for image caption Evaluation)	8
4.1.12.	REO (Relevance, Extraness, Omission)	8
4.1.13.	ViLBERTScore (Vision-and-Language BERT Score)	8
4.1.14.	LEIC (Learning to Evaluate Image Captioning)	9
4.1.15.	FAIER (Fidelity and Adequacy ensured Image caption Evaluation metric)	9
4.1.16.	NNEval (Neural Network based Evaluation Metric)	9
4.1.17.	LCEval (Learned Composite Metric for Caption Evaluation)	9
4.2.	Reference-free metrics	9
4.2.1.	CLIPScore (Contrastive Language–Image Pre-training Score)	9

* Corresponding author at: Federal Institute of Education, Science and Technology of Santa Catarina, Gaspar, 89111-009, SC, Brazil.

E-mail addresses: andrei.inacio@ifsc.edu.br (A.d.S. Inácio), hslopes@utfpr.edu.br (H.S. Lopes).

4.2.2.	UMIC (Unreferenced Metric for Image Captioning).....	9
4.2.3.	EMScore (Embedding Matching-based score)	10
4.3.	Timeline of automatic evaluation metrics	10
5.	Empirical experiments.....	11
5.1.	Popular metrics for video captioning.....	11
5.2.	Potential metrics for video captioning.....	11
5.3.	Specific metrics for video captioning.....	12
5.4.	Analysis.....	13
6.	Discussion.....	14
6.1.	Limitations of the evaluation metrics	14
6.2.	Possible extension to other languages.....	15
7.	Conclusions and research trends.....	15
	CRediT authorship contribution statement	16
	Declaration of competing interest.....	16
	Data availability	16
	Acknowledgments.....	16
	References.....	16

1. Introduction

In recent years, we have witnessed an exponential growth in the amount of images and videos produced and stored by people and enterprises and made available on the Internet. Understanding the visual content of images and videos and describing them in natural language has attracted the attention of researchers in the last few years (Aafaq, Mian, Liu, Gilani, & Shah, 2019; Rafiq, Rafiq, & Choi, 2021). Comparing images with videos, understanding the latter is much more challenging, since they require sophisticated techniques to process the diversity of human and object appearances that appear in diverse environments and also, with complex interactions between each other over time. An approach that accurately describes events in videos can be helpful in many applications, such as human–robot interaction, video indexing, assistance to the visually impaired, sign language understanding, and intelligent video surveillance, to name a few.

Assessing the quality of such systems is a complicated and subjective task. This happens because the captions, besides being grammatically well-formed and fluent, need to refer to the video properly (Stefanini et al., 2023). Human evaluation is the gold standard for assessing the quality of the captions. However, this is only sometimes possible since this task is too labor-intensive and inefficient (Bin, Shang, Peng, Ding, & Chua, 2021).

To circumvent this problem, some metrics have appeared over time for evaluating the quality of video captions. Four metrics, namely: BiLingual Evaluation Understudy (BLEU) (Papineni, Roukos, Ward, & Zhu, 2002), Consensus-based Image Description Evaluation (CIDEr) (Vedantam, Lawrence Zitnick, & Parikh, 2015), Metric for Evaluation of Translation with Explicit ORDERing (METEOR) (Banerjee & Lavie, 2005), and Recall-Oriented Understudy for Gisting Evaluation (ROUGE-L) (Lin, 2004), are the most frequently used for the evaluation of video captioning approaches. These metrics have become popular since they were used in the Large Scale Movie Description Challenge (LSMDC 2015) (Rohrbach et al., 2017) and the codes were released by the Microsoft Common Objects in COntext (MS COCO) Evaluation Server.¹ However, such metrics are sensitive to word overlap and fail to compare captions that convey almost the same meaning but describe the same content with no words in common. To assess the semantic content of captions, Semantic Propositional Image Caption Evaluation (SPICE) (Anderson, Fernando, Johnson, & Gould, 2016) was suggested. By processing them into a scene graph, it examines the candidate and reference sentences from the standpoint of their semantic similarity. Even so, since it only uses a dependency parser to analyze the sentences, it might be unable to interpret grammatical information.

The traditional evaluation metrics are based on n-gram overlapping. They basically compute the similarity of a candidate sentence against a set of reference sentences. Fig. 1 shows a straightforward example of evaluation using the conventional metrics to highlight the complexity of the challenge of automatically analyzing video captioning approaches, as well as the primary shortcomings presented in such metrics.

First, a reference sentence is provided (blue box) for each video. The next step is to assess the two hypothetical candidate sentences, A (red box) and B (green box). The first one is semantically correct, and the second one is wrong. The correct candidate sentence scored less than the incorrect one, according to the above-mentioned metrics. This is because they have fewer exact words in their reference sentences. Additionally, the accuracy of such measures is significantly hampered by the small number of reference phrases. Actually, the reference sentences required by these metrics may not completely cover the visual content because they are selective translations of the video made by human referees or an automated system (Jiang et al., 2020).

Unlike traditional metrics that require a set of reference sentences for evaluation, a recent promising metric called EMScore (Shi, Yang, Xu, Yuan, Li, Hu and Zha, 2021) was proposed to measure the similarity between a video and candidate sentences without using reference sentences during evaluation. Instead, it employs a large-scale vision-language model that was pre-trained to extract visual and linguistic features to compute a score based on the consistency of the video and caption. Using a pre-trained model reduces the gaps between the video and text embeddings. However, a significant semantic gap still exists between visual and language domains. The semantic gap (Baāzaoui, Barhoumi, Ahmed, & Zagrouba, 2018; Perlin & Lopes, 2015), can be understood as the “distance” between the low-level information (pixels, edges, shapes, texture) of images and their high-level meaning (language) in a given context.

Some metrics were developed from the Natural Language Processing (NLP) perspective. They are based on n-grams for estimating the semantic similarity between two blocks of text. Such a task is called Semantic Textual Similarity (STS) and usually outputs a percentage or ranking of similarity between texts (Chandrasekaran & Mago, 2021). One of the main challenges faced in such a scenario is the coexistence of many possible meanings for a word or phrase (polysemy) or the existence of two or more words having the same spelling or pronunciation but different meanings and origins (homonymy). For example, consider the sentences: “The man is cooking a dish” and “The man is washing the dish”. These two sentences contain the noun “dish”, which is an example of a polysemous word. Despite not being equivalent, such sentences achieve high scores on some traditional metrics because they have many words in common and have the same length. On the other hand, the sentences “The man is cooking dinner for his family” and “The man is preparing a meal for his loved ones” are equivalent and

¹ <https://github.com/tylin/coco-caption>

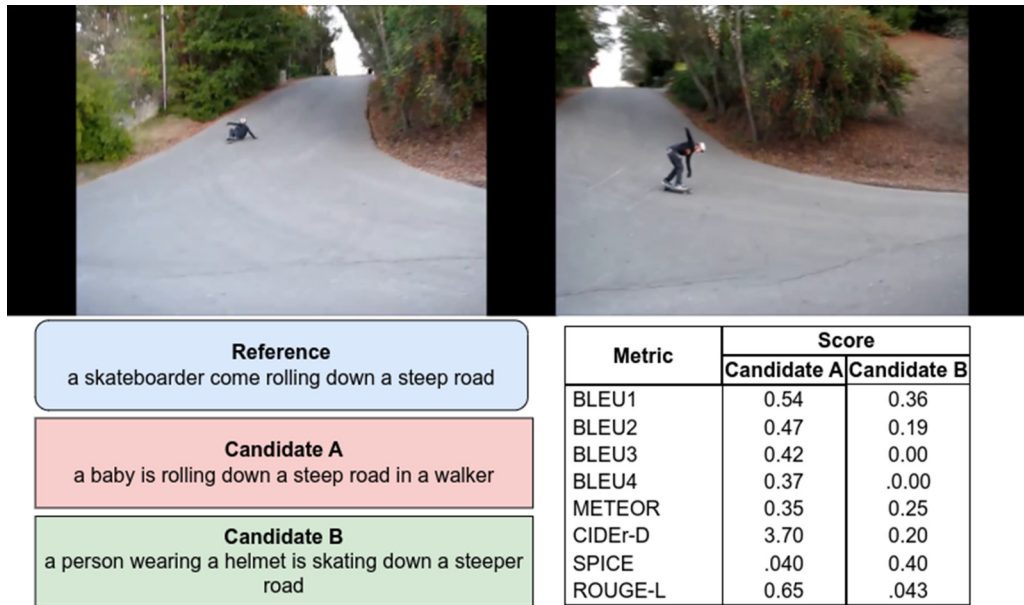


Fig. 1. An example of caption evaluation of a video from the ActivityNet Captions dataset (5pqVrMgiMcs) is shown. BLEU1, BLEU2, BLEU3, and BLEU4 denote the BLEU metric with 1-gram, 2, gram, 3-gram and 4-gram, respectively. Traditional reference-based metrics may fail to evaluate candidate sentences due to the small number of reference sentences. Moreover, despite not describing the video scene, candidate sentence “A” has more similar words to the reference sentence than candidate sentence “B”, and achieved a better score than the correct caption “A”. With the exception of CIDEr, which has a range of [0, 10], all metric scores are scaled in the range [0, 1]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

would achieve a lower score as they are written with different words. Besides, metrics that rely on visual information to compute the quality of a candidate sentence also face drawbacks due to the semantic gap problem.

Although some review studies on video captioning have been published in the last few years (Aafaq et al., 2019; Amirian, Rasheed, Taha, & Arabnia, 2020; Jain et al., 2022), they usually compare related methods, metrics, and datasets employed by existing approaches and disregard non-standard evaluation metrics. In a recent survey study on deep learning-based image captions (Stefanini et al., 2023), researchers analyzed non-standard metrics as an alternative or complement to standard metrics for more precise performance evaluation, even when ground-truth captions are not available during the inference step.

This paper presents a survey of the evaluation metrics for the video captioning task. To the best of our knowledge, this is the first in-depth review study about this subject. It is worth noticing that two popular metrics, CIDEr, and SPICE, were initially proposed for the image captioning task. Notwithstanding, they have been frequently used for evaluating video descriptions. Therefore, we also considered some promising metrics recently proposed for evaluating image captions in this survey. That said, the main contributions of this study are summarized as follows:

- A taxonomy of the existing metrics is proposed;
- The advantages and shortcomings of existing metrics are identified and discussed;
- The applications or contexts in which these metrics can be better used are suggested.
- An empirical comparison between the main metrics is shown to contrast their results.
- The main challenges in video captioning evaluation metrics are highlighted;

This paper is organized as follows. Section 2 presents a brief background and the challenges of the video captioning task. Section 3 presents the most popular video captioning benchmark datasets. In Section 4, the main evaluation metrics are overviewed. Section 5 presents empirical experiment results on videos from popular video captioning

datasets. Next, Section 6 discusses the limitations of the evaluation metrics. Finally, Section 7 presents the conclusions, and points out future research directions.

2. Video captioning vs image captioning

Humans can easily describe the visual content of images and videos using natural language. Notwithstanding, this is still a challenging task for computers. Generating natural language descriptions from visual content (images and videos) involves solving several complex problems, including: object detection and classification; human action recognition; detection of the visual relationships between humans and objects.

Image and video captioning tasks require the “translation” of visual content into a sequence of words, which can be seen as similar tasks. Instead of dealing with images with static structural information, the video captioning task has to process and understand the visual content presented in a sequence of frames, and translate them into a sequence of words. To achieve this, a video captioning approach must capture not only the individual frames but also their relationships and order in time. As a result, the approach must have a strong contextual understanding of the temporal content presented in the video. Moreover, the temporal component of videos introduces an additional level of difficulty, as it requires recognizing how the visual content evolves over time. This may involve tracking objects, detecting motion, and identifying actions. Thus, compared to image captioning, video captioning is more challenging as it requires sophisticated techniques to deal with the diversity of human and object appearances in different environments, as well as their changing relationships over time (Ji & Wang, 2021).

Nowadays, with the advancement of Computer Vision (CV) and Artificial Intelligence (AI) techniques, computers can effectively solve many real-world problems, including object classification, action recognition, and image segmentation. However, a step beyond the simple categorical classification of objects and actions is translating complex visual information into a semantically structured text (Inácio, Gutoski, Lazzaretti, & Lopes, 2021).

Early approaches proposed for video captioning started with template-based methods. In those approaches, the objects, activities, and scenes were first detected and then used in a sentence template (Aafaq et al., 2019; Liu, Xu, & Wang, 2019). Although these methods could generate descriptions based on grammar, they did not consider the spatial and temporal associations between entities. Inspired by the exponential development of deep learning techniques in the CV and NLP areas, video captioning research has recently emerged as a hot research topic. Usually, deep learning approaches are mainly designed as encoder–decoder pipelines. The encoder uses convolutional networks to convert the input visual content into a feature vector representation. The decoder is usually a Long-Short Term Memory (LSTM), Gated Recurrent Unit (GRU), or transformer-based network that outputs a sequence of words.

Once the captions are generated, an evaluation process is essential to assess the performance and effectiveness of the proposed systems and allow their comparison with other approaches. Human evaluation is often considered the gold standard, the ideal and most reliable metric to assess caption quality. However, it is always time-consuming, labor-intensive, and, sometimes, not consensual. Thus, automatic evaluation metrics are required to evaluate video captioning approaches.

3. Video captioning datasets

The evaluation metrics discussed in this work have been used to perform a quantitative analysis of video descriptions using datasets as benchmarks. Therefore, the metrics' performance is closely related to the quality, size, and diversity of the datasets. Existing video captioning models are trained on publicly available datasets and employ a hold-out validation strategy, following existing studies that use standard training, validation, and testing splits. This training approach ensures a fair comparison with state-of-the-art methods. The metrics presented in this study can be used to monitor model performance during training and to report performance on the test set post-training. To the best of our knowledge, no studies in the literature have employed a different validation strategy, possibly due to computational costs. Table 1 shows the main details of the most widely used datasets, which can be categorized into three domains: “open” (unspecific videos); “human” (focused on human-centered activities); and “cooking” (regarding cooking-related activities). Examples and detailed analysis of the datasets mentioned in Table 1, is outside the scope of this work, as well as papers where each metric was used, and can be found elsewhere (Aafaq et al., 2019; Amirian et al., 2020; Jain et al., 2022).

- ActivityNet Captions (Krishna, Hata, Ren, Fei-Fei, & Niebles, 2017): contains 20,000 videos taken from the ActivityNet dataset (Heilbron, Escorcia, Ghanem, & Niebles, 2015), in which each video has, on average, 3.65 temporally localized sentences and a total of 100,000 sentences. All videos were annotated by Amazon Mechanical Turk workers. The dataset was proposed for the dense video captioning task, which aims to generate multiple informative and diverse sentences for a video containing short, long, or even overlapping events.
- Charades (Sigurdsson et al., 2016): provides 27,847 descriptions of 9848 videos annotated by Amazon Mechanical Turk workers. Each video has an average length of 30 s and includes 15 types of indoor scenes of daily-life human activities. It is also available 66,500 temporally localized intervals for 157 action classes and 41,104 labels for 46 object classes. It was proposed for activity understanding, including action classification, localization, and video descriptions.
- Microsoft Research Video Description Corpus (MSVD) (Chen & Dolan, 2011): this is, possibly, the most used dataset for the video captioning task. It contains 70,028 sentences from 1970 video clips collected by Amazon Mechanical Turk workers. Each video contains one main activity to be described, usually lasting between 10 and 25 s. Moreover, the audio is muted in all video clips.

Table 1

Datasets used for the evaluation of video description approaches.

Dataset	Domain	#videos	#sentences	#vocabulary
ActivityNet Captions	Open	20,000	100,000	1,348,000
Charades	Human	9848	27,847	4144
MSR-VTT	Open	10,000	200,000	29,316
MSVD	Open	1970	70,028	13,010
TACoS	Cooking	14,105	52,593	2000
YouCook2	Cooking	2000	15,400	2600

- MSR Video to Text (MSR-VTT) (Xu, Mei, Yao, & Rui, 2016): contains 200,000 sentences for 10,000 clips extracted from 7180 videos, with an average of 20 different sentences per clip. All videos were annotated by Amazon Mechanical Turk workers. It is the second most used dataset.
- Saarbrücken Corpus of Textually Annotated Cooking Scenes datasets (short: TACoS) (Rohrbach et al., 2014): contains 52,593 descriptions of 14,105 video clips about people's cooking procedures. All descriptions were annotated by Amazon Mechanical Turk workers. It provides three levels of detailed descriptions for complex videos: one sentence for a complex event, a short sentence for a video segment, and a detailed description for each step of cooking procedures.
- YouCook2 (Zhou, Xu, & Corso, 2018): contains 15,400 sentences of video clips in 2000 untrimmed videos downloaded from YouTube, all instructional cooking recipe videos. The descriptions were provided by two human annotators. To date, this is the largest task-oriented instructional video dataset for the computer vision community.

4. Evaluation metrics

This Section reviews the automatic evaluation metrics commonly used for video captioning tasks. Moreover, we also consider in this study some metrics that were explicitly proposed for image captioning, but are useful and promising for the video captioning task too. We did not include studies that propose evaluation metrics for Natural Language Generation (NLG) systems, such as Machine Translation, Dialog Generation, Summarization, Question Answering, or other tasks different than video or image captioning.

We also propose a taxonomy that characterizes and classifies the automatic evaluation metrics based on their dependency on reference sentences, domain, and similarity aspects. A previous review study proposed a taxonomy for image captioning metrics (Sharif, Nadeem, Shah, Bennamoun, & Liu, 2020). In such a study, the metrics were divided into two categories: data-driven and hand-designed. Data-driven metrics involve learning to measure sentence correspondence through a data-driven approach, while hand-designed metrics use a set of hand-crafted criteria or features. Recently, many metrics have been proposed to assess captions directly from the visual content without reference sentences. Thus, the proposed taxonomy presented in this study differs from Sharif, Nadeem et al. (2020) by taking into account the unique features and aspects of the more recent metrics that are reported in this study.

An outline of a taxonomy for the metrics examined in this work is shown in Fig. 2. The metrics are divided into two primary categories: reference-based and reference-free. Reference-based metrics provide a similarity score between one or more reference sentences and a target sentence. Meanwhile, reference-free metrics score similarity between a target sentence and visual information (image or video). Then, each category can additionally be split into learned and hand-crafted subcategories. The hand-crafted approaches employ deterministic measures of similarity between a candidate and the reference sentences, such as the F-score or the cosine similarity. The learned methods usually require training a (neural network) model to predict the likelihood of a candidate caption being a human-generated description.

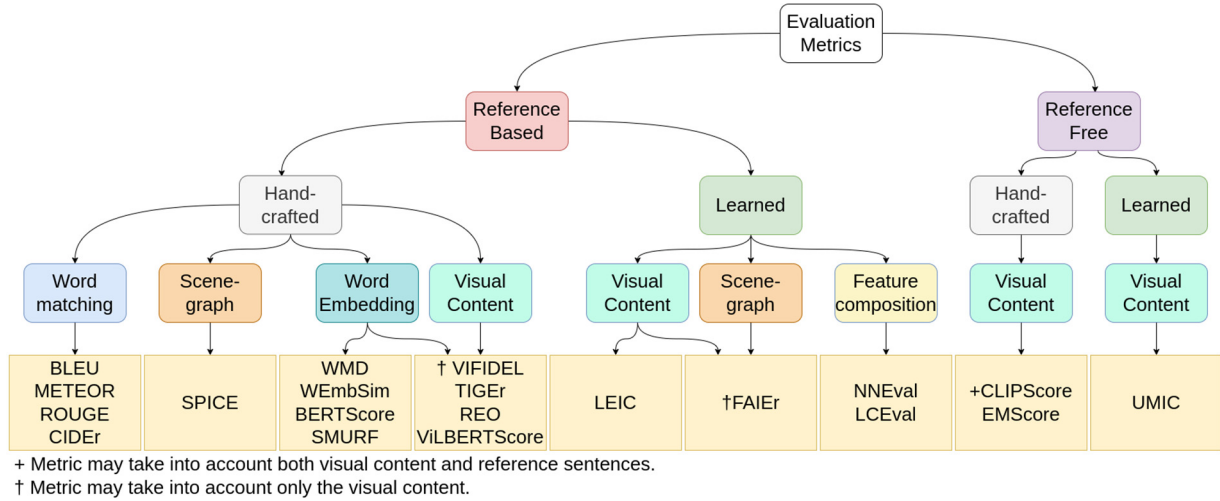


Fig. 2. Proposed taxonomy of the evaluation metrics.

In the proposed taxonomy, we also consider the way these metrics encode the sentences to compute the metrics, which can be divided into four main ways:

- Word-matching: when n-grams are compared;
- Scene-graph: when sentences are encoded as a scene-graph prior to comparison;
- Word embedding: when using a pre-trained word-embedding to encode sentences;
- Feature composition: when different features are considered.

Moreover, some metrics also include visual content (concepts captured in images) to measure the similarity. They were also categorized in the proposed taxonomy.

4.1. Reference-based metrics

Existing datasets for video captioning consist of a set of videos paired with captions in natural language, usually written by humans, which describe their visual content.

Most metrics used for evaluating video captioning approaches are based on those reference sentences. Thus, given a candidate sentence generated by the approach, the metrics evaluate the sentence by measuring its similarity against a set of reference sentences associated with a given visual content.

A brief description of the reference-based metrics is presented below. More details can be found in the original papers.

4.1.1. BLEU (BiLingual Evaluation Understudy)

BLEU (Papineni et al., 2002) is a quick, inexpensive, and language-independent method initially proposed for automatic evaluation of machine translation and is commonly used to evaluate image and video captioning approaches. It measures the overlapping precision of the n-grams of a predicted sentence with one or more reference human descriptions. BLEU is based on modified n-grams precision, and it is usually computed for n-grams of size 1 to 4. Grammatical correctness or intelligibility is not directly considered. A high score in this metric may be associated with a large number of references. BLEU scores range from 0 to 1 but are usually reported as a percentage value. A score above 0.30 generally reflects an understandable sentence, and above 0.50 reflects good and fluent candidate sentences (Denkowski & Lavie, 2010). The BLEU is computed as follows:

$$BP = \begin{cases} 1 & \text{if } c > r. \\ e^{(1-r/c)} & \text{if } c \leq r. \end{cases}$$

$$BLEU = BP \cdot \exp \left(\sum_{n=1}^N w_n \log p_n \right) \quad (1)$$

where, c denotes the length of the candidate sentence, r is the reference corpus length, BP is a brevity penalty factor to penalize candidate sentences shorter than reference sentences, w_n are positive weights summing to one, p_n is the geometric average of the modified n-gram precisions up to N , and \exp is the exponential function. Usually, N is set to 4 and w_n is set to $1/N$.

4.1.2. METEOR (Metric for Evaluation of Translation with Explicit Ordering)

METEOR (Banerjee & Lavie, 2005) is also a metric initially proposed for automatic machine translation, and it was designed to address the weakness perceived in the BLEU metric, including: the lack of recall, the use of higher-order n-grams, the lack of explicit word-matching between candidate and reference sentences, and the use of geometric averaging of n-grams. It creates an alignment between unigrams in the candidate and reference sentences. Each unigram from the candidate can have zero or one mapping to a unigram from the reference sentences. The METEOR word matching supports morphological variants, including stemming and synonyms. The metric is based on the precision, recall, and harmonic mean and consists of creating alignment between unigrams from candidate and reference sentences. The METEOR score is computed according to the following equation:

$$Score = Fmean(1 - Penalty) \quad (2)$$

The $Fmean$ is computed by combining Precision and Recall using the harmonic mean according to the following formula:

$$Fmean = \frac{10PR}{R + 9P} \quad (3)$$

where P and R stand for Precision and Recall and are computed as m/c and m/r , respectively, where m is the number of unigrams co-occurring in both candidate and reference sentences, c is the number of unigrams in the candidate sentence, and r is the number of unigrams in the reference sentence.

A penalty is calculated as follows to take into account the degree to which the corresponding unigrams in both the candidate and reference sentences are in the same word order.

$$Penalty = 0.5 \left(\frac{N_c}{N_u} \right)^3 \quad (4)$$

where the total number of matched unigrams is denoted by N_u , while N_c represents the smallest possible number of chunks, which are groups of matched unigrams that appear in the same order in both the candidate and reference sentences.

4.1.3. CIDEr (Consensus-based Image Description Evaluation)

CIDEr (Vedantam et al., 2015) is the first metric that was specifically proposed for evaluating image captioning approaches. It proposes a consensus-based evaluation protocol using the Term Frequency-Inverse Document Frequency (TF-IDF) to capture the frequency of each word in a candidate sentence in a list of reference sentences. The main idea is to evaluate how well a candidate sentence c_i matches the consensus of a set of image descriptions $S_i = \{s_{i1}, s_{i2}, \dots, s_{im}\}$. Each sentence s_{ij} is represented as a set of n -grams, and a given n -gram w_k is a set of one or more words. TF-IDF $g_k(s_{ij})$ for each n -gram w_k is computed using:

$$g_k(s_{ij}) = \frac{h_k(s_{ij})}{\sum_{w_l \in \Omega} h_l(s_{ij})} \log \left(\frac{|I|}{\sum_{I_p \in I} \min(1, \sum_q h_k(s_{pq}))} \right) \quad (5)$$

Where, $h_k(s_{ij})$ is the number of times an n -gram w_k occurs in a reference sentence s_{ij} , $h_k(c_i)$ is the number of times an n -gram w_k occurs in a candidate sentence c_i , Ω is the vocabulary of all n -grams and I is the number of all images in the dataset. The TF term confers greater weighting to n -grams that exhibit higher frequency in the reference sentence utilized for image description, while the second term of $g_k(s_{ij})$, IDF, attenuates the weighting of n -grams that exhibit frequent occurrence across all images in the dataset by dividing the number of images in which w_k appears in any of its reference captions.

The, the similarity between each reference caption s_{ij} and a candidate sentence c_i is computed by the average cosine distance of the TF-IDF vectors.

$$CIDEr_n(c_i, S_i) = \frac{1}{m} \sum_j \frac{g^n(c_i)g^n(s_{ij})}{\|g^n(c_i)\| \|g^n(s_{ij})\|} \quad (6)$$

where, $g^n(c_i)$ is a vector consisting of all n -grams $g_k(c_i)$ of length n and $\|g^n(c_i)\|$ is the magnitude of the vector $g^n(c_i)$. The same definition is used for $g^n(s_{ij})$.

When using longer n -grams, it is possible to capture rich semantic information and grammatical properties. The CIDEr with multiple lengths of n -grams can be calculated as:

$$CIDEr(c_i, S_i) = \sum_{n=1}^N w_n CIDEr_n(c_i, S_i) \quad (7)$$

where w_n was empirically define by the author as $1/N$.

The CIDEr-D is a variation of the CIDEr and is commonly used to evaluate image and video captioning applications. It introduces a Gaussian penalty based on the difference between candidate and reference sentence lengths. Moreover, a clip to the n -gram counts in the CIDEr numerator is considered. These modifications aim to avoid sentences with high scores but with poor results when judged by humans. The CIDEr-D score is defined as follows:

$$CIDEr-D(c_i, S_i) = \sum_{n=1}^N w_n CIDEr-D_n(c_i, S_i) \quad (8)$$

$$CIDEr-D_n(c_i, S_i) = \frac{10}{m} \sum_j e^{-\frac{(l(c_i) - l(s_{ij}))^2}{2\sigma^2}} * \frac{\min(g^n(c_i), g^n(s_{ij})) \cdot g^n(s_{ij})}{\|g^n(c_i)\| \|g^n(s_{ij})\|} \quad (9)$$

where, c_i and $S_i = \{s_{i1}, \dots, s_{im}\}$ are a candidate sentence and a set of m reference sentences for an image i , $w_n = 1/N$ and $N = 4$ are uniform weight and n -gram order defined empirically by the authors, $l(c_i)$ and $l(s_{ij})$ denotes the lengths of candidate c_i and reference sentence s_{ij} , respectively. The authors also defined $\sigma = 6$, and to ensure that the CIDEr-D scores are comparable to other metrics, a factor of 10 was added.

4.1.4. ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

This package, developed by Lin (2004), was aimed at the automatic evaluation of summaries. It consists of four different metric variations: ROUGE-N (N-gram Co-Occurrence Statistics), ROUGE-L (Longest

Common Subsequence), ROUGE-W (Weighted Longest Common Subsequence), and ROUGE-S (Skip-Bigram Co-Occurrence Statistics). The ROUGE-L metric is often used to evaluate image and video captioning approaches. It is a recall-based approach that uses the F-measure to compute the score, using n -gram overlapping and longest common subsequences between two statements. The ROUGE-L is computed by the following equations:

$$R_{lcs} = \frac{LCS(X, Y)}{|X|} \quad (10)$$

$$P_{lcs} = \frac{LCS(X, Y)}{|Y|} \quad (11)$$

$$ROUGE_L = F_{lcs} = \frac{(1 + \beta^2)R_{lcs}P_{lcs}}{R_{lcs} + \beta^2P_{lcs}} \quad (12)$$

where, $LCS(X, Y)$ denotes the length of a longest common subsequence of X and Y , $|X|$ is the length of X , $|Y|$ is the length of Y , and β controls the relative importance of P_{lcs} and R_{lcs} and is usually set to 1.2 (Chen et al., 2015).

4.1.5. SPICE (Semantic Propositional Image Caption Evaluation)

This metric was designed by Anderson et al. (2016) to tackle the limitations of the existing automatic evaluation metrics based on n -grams, such as BLEU, METEOR, and CIDEr. Usually, these metrics assign a low score to a generated sentence that conveys almost the same reference meaning but has no words in common.

It was originally proposed for the image captioning task, but it is also employed to evaluate video captioning systems. The metric encodes objects, attributes, and relations from candidate and reference sentences in graph-based semantic representations $G(c)$ and $G(S)$, respectively, by using a dependency parse tree.

$$G(c) = \langle O(c), E(c), K(c) \rangle \quad (13)$$

where, $O(c) \subseteq C$ is a set of objects mentions in a sentence c , $E(c) \subseteq O(c) \times R \times O(c)$ is the set of hyper-edges representing relations between objects, and $K(c) \subseteq O(c) \times A$ is the set of attributes associated with objects.

During the match analysis between tuples, synonym and lemmatization techniques are considered to allow the match of words with different inflection forms. The logical tuples from a scene graph is defined the function T , as:

$$T(G(c)) \triangleq O(c) \cup E(c) \cup K(c) \quad (14)$$

Each tuple may contain one, two, or three elements, representing objects, relations and attributes, respectively. The caption quality is calculated based on the F1-score over tuples in the candidate and reference sentences, and can be defined as:

$$P(c, S) = \frac{|T(G(c)) \otimes T(G(S))|}{|T(G(c))|} \quad (15)$$

$$R(c, S) = \frac{|T(G(c)) \otimes T(G(S))|}{|T(G(S))|} \quad (16)$$

$$SPICE(c, S) = F_1(c, S) = \frac{2P(c, S)R(c, S)}{P(c, S) + R(c, S)} \quad (17)$$

where, \otimes is a binary matching operator that returns matching tuples in two scene graphs.

4.1.6. WMD (Word Mover's Distance)

It is a distance measure proposed by Kilickaya, Erdem, Ikizler-Cinbis, and Erdem (2017) to calculate the dissimilarity between two text documents. It was inspired by the "Earth Mover's Distance" (EMD), employing a solver of the "transportation problem".

This metric aimed to assess the semantic distance between documents by representing the words as word embedding vectors. It calculates the minimum distance that words in one document should travel to the words in another document.

This metric was not designed for image or video captioning evaluation. However, it has been used to evaluate image captioning approaches (Laina, Rupprecht, & Navab, 2019) and, over time, it has inspired the development of other metrics.

The resulting WMD score represents the dissimilarity or distance between the two documents, with a lower value indicating higher similarity and a higher value indicating lower similarity. The WMD distance between documents x and y is defined as:

$$\min_{T \geq 0} \sum_{i,j=1}^n T_{ij} c(i,j)$$

$$\text{subject to: } \begin{cases} \sum_{j=1}^n T_{ij} = d_i \quad \forall i \in \{1, \dots, n\} \\ \sum_{i=1}^n T_{ij} = d'_j \quad \forall j \in \{1, \dots, n\} \end{cases} \quad (18)$$

where T is a flow matrix in $\mathbb{R}^{(n \times n)}$ and $T_{ij} \geq 0$ means how much of word i in one document travels to word j in another document, n is the vocabulary size, $c(i,j) = \|x_i - x_j\|_2$ is the distance between word i and word j encoded in m -dimensional embedding space, d_i is the word frequency that appears c_i times in the document, d'_j is the word frequency that appears c_j times in the document.

4.1.7. WEmbSim

Similar to WMD, WEmbSim (Sharif, White, Bennamoun, Liu, & Shah, 2020) uses word embeddings to encode the words in an embedding space. Using an embedding matrix V , each sentence is mapped to a vector representation via the Mean of Word Embeddings (MOWE), as defined in Eq. (19) and denoted by the function $\tilde{v}(\cdot)$. Then, the distance between two sentences is computed by the cosine similarity (cossim), as follows.

$$\tilde{v}(C) = \frac{1}{n} \sum_{w_i \in C} V_{\cdot, w_i} \quad (19)$$

$$\text{cossim}(\tilde{a}, \tilde{b}) = \frac{|\tilde{a} \cdot \tilde{b}|}{|\tilde{a}| |\tilde{b}|} \quad (20)$$

$$\text{Score}(C|R) = \Phi \text{cossim}(\tilde{v}(C), \tilde{v}(R_i)), \forall R_i \in R \quad (21)$$

where, Φ is a rule used to specify how to combine the score for multiple reference sentences. The authors suggest using the *mean* combination function, as it consistently shows better performance than the *min* or *max* rule combination function, $\tilde{v}(\cdot)$ is a function which maps a given candidate sentence $C = [w_1, w_2, \dots, w_n]$ or a reference sentence $R_i = [w_1, w_2, \dots, w_n]$ into a feature vector representation, n is the number of words in a given sentence, and i is the index of the i th reference sentence.

WEmbSim was developed as an automatic evaluation metric for image captioning systems, measuring system-level performance based on semantic similarity. However, similar to SPICE, it does not consider fluency and may struggle to distinguish between sentences with the same words in different orders.

4.1.8. BERTScore (Bidirectional Encoder Representations from Transformers Score)

BERTScore (Zhang, Kishore, Wu, Weinberger, & Artzi, 2020) is an automatic metric for machine translation and image captioning systems. It uses the BERT model (Devlin, Chang, Lee, & Toutanova, 2019) to extract token level vector representation from candidate c and reference r sentences. Then, the Precision and Recall metrics are computed as follows:

$$P_{BERT} = \frac{1}{|c|} \sum_{c_j \in c} \max_{r_i \in r} (r_i^T c_j) \quad (22)$$

$$R_{BERT} = \frac{1}{|r|} \sum_{r_i \in r} \max_{c_j \in c} (r_i^T c_j) \quad (23)$$

$$F_{BERT} = 2 \frac{P_{BERT} R_{BERT}}{P_{BERT} + R_{BERT}} \quad (24)$$

where r_i and c_j are token level vector representations from r and c sentences, respectively.

Similar to cosine similarity, the final score of BERTScore metric ranges between -1 and 1 . Higher BERTScore values indicate better similarity between the generated and reference text, while lower values indicate lower similarity. However, the scores are often in the upper end of that range. Thus, the authors suggest the use of a baseline scaling to increase the score readability, leaving the final score in the range $[0, 1]$. The rescaling procedure for R_{BERT} is:

$$R_{SBERT} = \frac{R_{BERT} - b}{1 - b} \quad (25)$$

where, b is an empirical lower bound on observed BERTScore. The same rescaling procedure have to be applied for P_{BERT} and F_{BERT} .

4.1.9. SMURF (SeMantic and linguistic UndeRstanding Fusion)

SMURF (Feinglass & Yang, 2021) is an automatic evaluation metric that combines a novel semantic evaluation algorithm SPARCS (Semantic Proposal A likeness Rating using Concept Similarity) and novel fluency evaluation algorithms SPURTS (Stochastic Process Understanding Rating using Typical Sets) and MIMA (Model-Integrated Meta-Analysis) for both caption-level and system-level analysis. A Transformer-based model such as BERT or RoBERTa (Liu, Ott et al., 2019) is used to extract features from texts and capture both the syntax and morphology of the text.

MIMA was proposed to estimate the typicality into evaluation of a candidate sentence y^n as follows.

$$f_{MIMA}(y^n, p) = 1 - \text{median}_{\text{layer}}(\max_{\text{head}}[I_{\text{flow}}(y^n, p)]) \quad (26)$$

$$I_{\text{flow}}(y^n, p) = MI = \frac{2H(\alpha_i(y^n, p)) + H(\alpha_j(y^n, p)) - H(\alpha_{ij}(y^n, p))}{H(\alpha_i(y^n, p)) + H(\alpha_j(y^n, p))} \quad (27)$$

where I_{flow} denotes the information flow in terms of the attention dimensions $\alpha_i(y^n, p)$, $\alpha_j(y^n, p)$, and their joint distribution $\alpha_{ij}(y^n, p)$, MI is the normalized mutual information, defined in Witten and Frank (2005), which is a measure of the mutual dependence or redundancy between two sets of random variables, α_{ij} are attention layers weights computed by the distilled BERT model from a sequence vector of tokenized words of a candidate sentence y^n , $H(\alpha_i(y^n, p))$ is the entropy of the attention distribution $\alpha_i(y^n, p)$ for the i th attention dimension, $H(\alpha_j(y^n, p))$ is the entropy of the attention distribution $\alpha_j(y^n, p)$ for the j th attention dimension, $H(\alpha_{ij}(y^n, p))$ is the entropy of the joint attention distribution $\alpha_{ij}(y^n, p)$ between the i th and j th attention dimensions.

MIMA serves as a basis for evaluating the fluency of input text, which can be divided in grammar and style. Grammar depends of the typicality of the whole sentence and is computed using the f_{MIMA} . Style depends on the distinctness or atypicality of the words directly associated with the image description. Thus, focusing on style, SPURTS was proposed to evaluate the distinctness or atypicality of the words in the candidate sequence without stop words (denoted as $y_{w/o}$). Here, the distilled RoBERTa model was used since it performs well on out-of-distribution.

$$\text{SPURTS} = 1 - f_{MIMA}(y_{w/o}, p) \quad (28)$$

The SPARCS metric mainly focuses on semantics and is defined as follows.

$$P(C, S) = \frac{\sum_i \frac{df_{gt(S)}(C_i)}{|gt(S)|}}{\sum_i \left(\frac{df_{gt(S)}(C_i)}{|gt(S)|} + \mathbb{I}[df_{gt(S)}(C_i) = 0] \right)} \quad (29)$$

$$R(C, S) = \frac{\sum_i df_{gt(S)}(C_i)}{\sum_i df_{gt(S)}(S_i)} \quad (30)$$

$$\text{SPARCS} = F_1(C, S) = \frac{2P(C, S)R(C, S)}{P(C, S) + R(C, S)} \quad (31)$$

Where C is the candidate concept set, $gt(S)$ is the reference caption set, gt is a function that maps concepts to a reference caption set, and df is the document frequency, which is used to estimate typicality of the concept across the sentences.

Finally, the SMURF metric can be defined as follow.

$$SMURF = \begin{cases} SPARCS' + G & \text{if } SPARCS' < T, \\ SPARCS' + D + G & \text{otherwise} \end{cases}$$

Where $G = \min(MIMA' - T, 0)$ is a grammatical outlier penalty, $D = \max(SPURTS' - T, 0)$ is a style reward, and $T = -1.96$ is an empirically threshold defined by the authors.

4.1.10. VIFIDEL (Visual Fidelity for Image Description Evaluation)

VIFIDEL (Madhyastha, Wang, & Specia, 2019) was developed inspired on the WMD metric to estimate the faithfulness of a generated caption concerning the content of a given image. It measures the similarity between objects detected in the image and the words in the generated caption using the WMD metric. Additionally, it can incorporate reference descriptions when available to enhance the evaluation.

$$VIFIDEL(I, S) = \exp(-WMD(d^I, d^S)) \quad (32)$$

where d^I is a semantic vector representation containing normalized bag of object category labels for image I and d^S is the normalized bag of words representation for description S .

This metric can be extended to utilize sentence references, when available, to assess the importance of objects in an image. Let $R^I = (R_1^I, R_2^I, \dots, R_M^I)$ be a set of human references for a given image I , a penalty weight p_j^k , for a word k (object label in image I or a word in a candidate sentence S^I) is computed as:

$$p_k^I = \frac{1}{M} \sum_{r=1}^M \left(\frac{1 - \max_{i \in \{R_r^I\}} \cos(x_k, x_i)}{2} \right) \quad (33)$$

$$c^I(i, j | R^I) = \|p_i^I x_i - p_j^I x_j\|_2^2 \quad (34)$$

where $\{R_r^I\}$ is the set of content words in the r th reference for image I , and x_i is the word embedding for word i . Replacing the cost $c(i, j)$ of WMD (see Eq. (18)) with Eq. (34), VIFIDEL score is computed considering a score weighted by object importance.

4.1.11. TIGER (The Text-to-Image Grounding based metric for image caption Evaluation)

The TIGER metric (Jiang et al., 2020) has been proposed for evaluating image captioning systems, taking into account both the image content and sentence references. To compute features from an image-sentence pair in a common semantic space, the metric uses the pre-trained Stacked Cross Attention Neural Network (SCAN) (Lee, Chen, Hua, Hu, & He, 2018), which is an image-text grounding model. The captions are encoded in a sequence of d -dimensional vectors, and the images are encoded in a set of $n = 36$ region-level 2048-dimensional features. The quality of a candidate sentence C based on a set of reference sentences R and a image V is then computed by combining two metric systems: Region Rank Similarity (RRS) and Weight Distribution Similarity (WDS).

$$RSS_{(V,C,R)} = \frac{DGC_{s(V,C)}}{IDCG_{s(V,R)}} \quad (35)$$

$$WDS_{(V,C,R)} = 1 - \frac{\exp(\tau D(R|C))}{\exp(\tau D(R|C)) + 1} \quad (36)$$

$$TIGER_{(V,C,R)} = \frac{RSS_{(V,C,R)} + WDS_{(V,C,R)}}{2} \quad (37)$$

where $s(V, C) = \{s_1, s_2, \dots, s_n\}$ is a set of similarity score between a candidate sentence C and all image regions, which can be ranked. $DGC_{s(V,C)}$ is based on the Discounted Cumulative Gain Järvelin and Kekäläinen (2002), which is used to measure quality of document

ranking in web search engines. Similarly, the $IDGC_{s(V,R)}$ is the Ideal DGC computed based on the reference sentences. $WDS_{(V,C,R)}$ is based on KL Divergence (Kullback & Leibler, 1951) and measures the distance between the two distributions. The final score ranges from 0 to 1, where a higher score indicates a better caption.

4.1.12. REO (Relevance, Extraneous, Omission)

REO metric (Jiang et al., 2019) provides a more informative assessment compared to other metrics as it generates scores from three different perspectives: Relevance, Extraneous, and Omission. To extract features from images and sentences (references and candidate), REO also employs the SCAN model, which creates a multimodal semantic space. The relevance score is then calculated using the cosine similarity (cossim) distance between the candidate and reference features, as shown below:

$$R = \frac{1}{N} \sum_{i=1}^N \text{cossim}(a_i^C, g_i) \quad (38)$$

where, a_i^C is the context features of the candidate sentence and g_i denotes either image features or context features extracted from reference sentences.

Extraneous scores are calculated by computing the similarity distance between the vertical context vector $a_{i\perp}^C$ and its original context vector a_i^C , as follows:

$$a_{i\perp}^C = a_i^C - \frac{a_i^C g_i}{\|g_i\|^2} g_i \quad (39)$$

$$E = \frac{1}{N} \sum_{i=1}^N d(a_i^C, a_{i\perp}^C) \quad (40)$$

where a_i^C is the context features of the candidate sentence, $a_{i\perp}^C$ represents the irrelevant content of C to the ground truth at i_{th} image region, and d is the Mahalanobis distance.

Similar to Extraneous, the Omission score is calculated as follows:

$$g_{i\perp} = g_i - \frac{g_i a_i^C}{\|a_i^C\|^2} a_i^C \quad (41)$$

$$O = \frac{1}{N} \sum_{i=1}^N d(g_i, g_{i\perp}) \quad (42)$$

where $g_{i\perp}$ represents the vertical context features based on the orthogonal projection of g_i to a_i^C .

4.1.13. ViBERTScore (Vision-and-Language BERT Score)

Inspired by the excellent performance of word-embedding techniques, especially the BERTScore model, in many text generation tasks, ViBERTScore (Lee et al., 2020) was proposed. It computes image-conditioned embeddings for each token using ViBERT (Lu, Batra, Parikh, & Lee, 2019) from both generated and reference texts. A cosine similarity among the pair of tokens from the candidate and reference caption is computed. The greedy matching process between these tokens is expressed via the cosine similarity of their embeddings. The best matching token pairs are used for computing precision, recall, and F1-score, as follows.

$$\text{ViBERTScore}_P = \frac{\sum_{i=1}^m \max_{\hat{h}_{w_j} \in H_{\hat{X}V}} h_{w_i}^T \hat{h}_{w_j}}{m} \quad (43)$$

$$\text{ViBERTScore}_R = \frac{\sum_{i=1}^n \max_{h_{w_j} \in H_{XV}} \hat{h}_{w_i}^T h_{w_j}}{n} \quad (44)$$

$$\text{ViBERTScore}_F = 2 \frac{\text{ViBERTScore}_P \text{ViBERTScore}_R}{\text{ViBERTScore}_P + \text{ViBERTScore}_R} \quad (45)$$

where $H_{XV} = (h_{w_0}, \dots, h_{w_T})$ and $H_{\hat{X}V} = (\hat{h}_{w_0}, \dots, \hat{h}_{w_T})$ are contextual embeddings provided from the pre-trained ViBERT for reference and candidate sentences, respectively. Note that ViBERT model compute features from a pair of image and caption embeddings.

4.1.14. LEIC (Learning to Evaluate Image Captioning)

The LEIC metric (Cui, Yang, Veit, Huang, & Belongie, 2018) is a discriminative evaluation technique that relies on machine learning to distinguish between human-written and machine-generated captions. It encodes the candidate and reference captions (when available) and images as feature vectors, which are then used as input into a softmax classifier to obtain the probability of the description being generated by a human or a machine, as follows.

$$score_{\theta}(\hat{c}, i) = P(\hat{c} \text{ is human written } | C(i), \theta) \quad (46)$$

where \hat{c} is the candidate sentence, $C(i)$ is the context of image i , which can include the reference caption as part of context, and θ is a learned parameter. Further information regarding the training and inference procedures can be found in the original paper.

4.1.15. FAIEr (Fidelity and Adequacy ensured Image caption Evaluation metric)

FAIEr (Wang, Yao, Wang, Wu, & Chen, 2021) is a learning-based metric that evaluates the fidelity and adequacy of captions generated by image captioning systems. It employs the same scene graph parser used by the SPICE metric to represent sentences as textual scene graphs. To create a visual scene graph, an object detector is employed to detect and extract object features from an image. Each detected object is a graph node, and the relationship-level representation is encoded using a Graph Convolutional Network (GCN). The visual and reference scene graphs are fused using an attention mechanism. The final score is computed by measuring the similarity between two scene graphs at the object and relationship levels.

$$S_o = \frac{\sum_{k=1}^{L_c^o} \max_{i \in [1, N_o]} (z_i^{oT} h_{ck}^o)}{L_c^o} \quad (47)$$

$$S_r = \frac{\sum_{k=1}^{L_c^r} \max_{i \in [1, N_o]} (z_i^{rT} h_{ck}^r)}{L_c^r} \quad (48)$$

where z_i^o and z_i^r are the union of object-level and relationship-level vector representations computed by the fusion of visual and reference scene graphs, h_{ck}^o and h_{ck}^r are object-level vector representations of candidate sentence and reference sentence. L_c^o and L_c^r are number of in candidate and reference sentences, respectively. The final score of the candidate caption with respect to the union reference information is $S = S^o + S^r$.

4.1.16. NNEval (Neural Network based Evaluation Metric)

NNEval (Sharif, White, Bennamoun, & Shah, 2018) is also a learning-based metric designed to evaluate image captioning system. It leverages both lexical and semantic information by using a composition of well-established output metrics such as BLEU, METEOR, CIDER, SPICE, and WMD. Rather directly using candidate and reference sentences to train the metric, NNEval uses a set of composed features derived from the scores generated by each individual metric. Then, the feature vector is used to feed a feed-forward neural network, that computes the probability of an input sentence being human-generated. The output can be formulated as follows:

$$P(k = 1, x) = \frac{e^{z_1}}{e^{z_0} + e^{z_1}} \quad (49)$$

where z_k represents unnormalized class scores (z_0 and z_1 correspond to the machine and human class respectively), and $x = \{x_1, x_2, \dots, x_i\}$ is a fixed length composite feature vector. More information about the network architecture as well as the training and inference processes can be found in the original paper.

4.1.17. LCEval (Learned Composite Metric for Caption Evaluation)

LCEval (Sharif, White, Bennamoun, Liu, & Shah, 2019) is a learning-based metric that extends the NNEval metric by incorporating various computed metrics. However, unlike NNEval, which combines all features into a feature vector, LCEval divides the features into three subgroups based on their lexical, semantic, and syntactic properties. The lexical features include BLEU, METEOR, ROUGE-L, and CIDER scores. The semantic features consider SPICE, WMD, and MOWE scores. Finally, the syntactic features are extracted using the Head Word Chain Matches (HWCM), which captures the syntactic similarity between sentences using the tree structure of the sentences. The final score can be formulated using Eq. (49).

4.2. Reference-free metrics

Due to the known limitations of the existing metrics based on reference sentences, mainly regarding the difficulty of obtaining several possible ways of describing the same visual content, some reference-free metrics were recently proposed. In such metrics, visual and textual features are extracted using pre-trained neural network models for the image-text matching task. Then, a similarity score is computed. A brief description of the reference-free metrics studied is described below. More information about the training and inference processes, as well as the architectures of the following metrics can be found in the original papers.

4.2.1. CLIPScore (Contrastive Language-Image Pre-training Score)

This metric was introduced by Hessel, Holtzman, Forbes, Le Bras, and Choi (2021) for assessing image captioning systems without reference sentences. It uses the CLIP (Radford et al., 2021) model, a cross-modal retrieval model pre-trained on 400M image + caption pairs, to extract features from images and candidate sentences. The final score is then computed by measuring the cosine similarity between features. Additionally, the metric can be extended to incorporate reference sentences when available. Given an image with visual CLIP embedding v and a candidate sentence with textual CLIP embedding c , the CLIPScore can be computed as follows:

$$CLIP-S(c, v) = \max(\cos(c, v), 0)w \quad (50)$$

where w was empirically defined by the authors as 2.5. To compute corpus-level CLIP-S, the average over the pairs (image, candidate) can be performed.

When reference sentences are available, the CLIPScore can be calculated as follows:

$$RefCLIP-S(c, R, v) = H-Mean(CLIP-S(c, R, v), \max_{r \in R} (\max \cos(c, r), 0)) \quad (51)$$

where R denotes the set of textual CLIP embedding references, and $H-Mean$ denotes the harmonic mean.

4.2.2. UMIC (Unreferenced Metric for Image Captioning)

UMIC (Lee, Yoon, Derroncourt, Bui, & Jung, 2021) is another free-reference metric designed to evaluate the quality of sentences generated by image captioning systems. It uses image features extracted from the UNITER (UNiversal Image-Text Representation learning) (Chen et al., 2020), a pre-trained model for predicting alignment between images and texts. The model is fine-tuned via contrastive learning to distinguish between the reference sentences and negative captions using synthetic negative samples. The UMIC score can be formulated as follows:

$$S(I, X) = \text{sigmoid}(W_{i_{[CLS]}} + b), \quad (52)$$

where $i_{[CLS]}$ is a joint representation of the input image and input caption computed by the UNITER, W and b are trainable parameters.

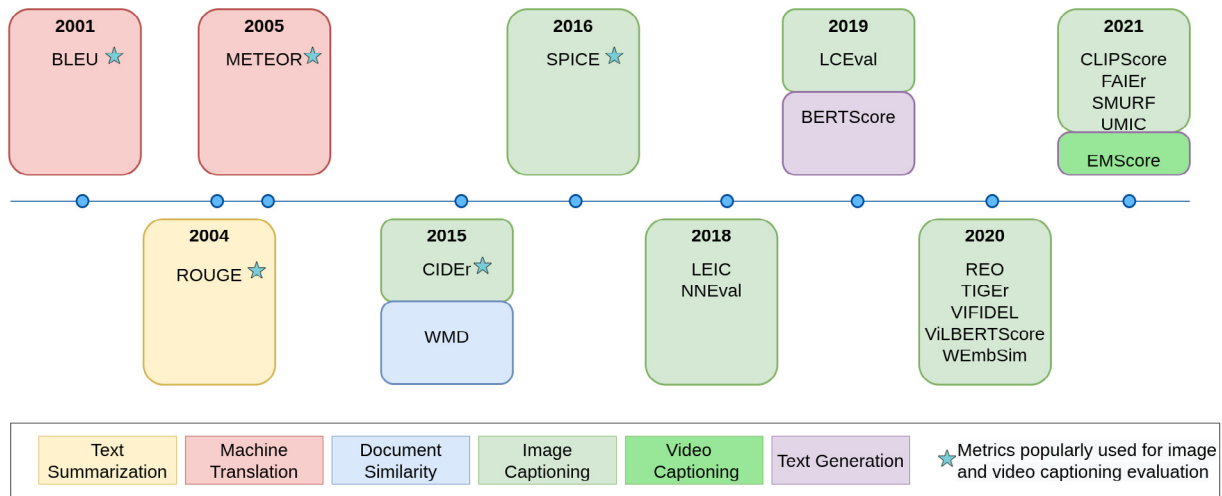


Fig. 3. Timeline of metrics classified by tasks: Text Summarization, Machine Translation, Document Similarity, Image Captioning, Video Captioning, and Text Generation.

4.2.3. EMScore (Embedding Matching-based score)

EMScore (Shi et al., 2022) is a free-reference metric proposed for evaluating video captioning approaches. It uses the pre-trained image-language model CLIP to extract video and text embeddings. To provide a comprehensive comparison between the video and caption, EMScore calculates the average matching scores at both the coarse-grained level (based on the global embeddings of the video and the candidate caption) and the fine-grained level (based on the embedding similarities between the frames and words).

For course-grained embedding matching, the score is computed using the following equation:

$$EMScore(X, V)_c = f_X^T f_V \quad (53)$$

where f_V and f_X are embeddings of the video and the captions, respectively. As all the embeddings are normalized using L2 normalization, the cosine similarity can be simplified to the inner product.

For fine-grained embedding matching, the score is computed using the precision (P), recall (R), and the F1 score, as follows:

$$P(X, V)_f = \frac{1}{|X|} \sum_{x_i \in X} \max_{v_j \in V} f_{x_i}^T f_{v_j} \quad (54)$$

$$R(X, V)_f = \frac{1}{|V|} \sum_{v_j \in V} \max_{x_i \in X} f_{x_i}^T f_{v_j} \quad (55)$$

$$EMScore(X, V)_f = 2 \frac{PR}{P+R} \quad (56)$$

where f_{x_i} and f_{v_j} are embeddings of the caption and the frame, respectively, $|X|$ is the number of tokens of a candidate sentence X , $|V|$ is the number of frames of a video V .

The final score is computed by combining a fine-grained score and a coarse-grained score as follows:

$$EMScore(X, V) = \frac{EMScore(X, V)_c + EMScore(X, V)_f}{2} \quad (57)$$

When reference sentences are available, they can also be considered as an extended metric called EMScore_ref, defined as follow.

$$EMScore(X, V, X^*)_{ref} = \frac{EMScore(X, V) + EMScore(X, X^*)}{2} \quad (58)$$

where V is a video content, X is the candidate sentence, and X^* is reference sentence. When dealing with multiple reference sentences $\{X_i^*\}_{i=1}^M$, $EMScore(X, X^*) = \max EMScore(X, X_i^*)$

4.3. Timeline of automatic evaluation metrics

Fig. 3 shows the timeline of the above-mentioned evaluation metrics. The timeline presents the metrics proposed for the image or video

captioning task, as well as those initially proposed for other tasks, but also used to report the performance of visual description systems. We categorized them in different colors, each representing a task for which they were primarily devised. Also, the popular metrics used for video captioning were highlighted with a star to emphasize that most of them were proposed for a different task than video captioning.

In the early approaches, BLEU, METEOR, ROUGE-L, and CIDEr were employed to evaluate video captioning (Venugopalan et al., 2015) using the code available in GitHub.² Later, SPICE was included in the library. It is a metric specifically proposed to assess the propositional semantic content from image captioning. Since then, these five metrics have become a kind of standard for reporting the state-of-the-art performance of video and image captioning approaches.

In 2015, the WMD metric was introduced for the document similarity task. It uses word embeddings to calculate the similarity between documents. Despite not being used directly for the video description task, it served as inspiration for other metrics proposed later, including VIFIDEL, WEmbSim, and BERTScore. WEmbSim and VIFIDEL use pre-trained word embeddings, such as word2vec, GLOVE, or fasttext. On the other hand, BERTScore is based on the BERT Model and was proposed for text generation and image captioning using contextualized embeddings.

Learned metrics, such as LEIC, NNEval, and LCEval, also have been proposed to improve evaluations at the caption level. Both NNEval and LCEval cast the problem of evaluation as a classification task. They consist of training a multi-layer feedforward neural network using different metrics scores as input, including BLEU, CIDEr, SPICE, and WMC, to distinguish between human and machine-generated captions. LEIC uses both the reference sentences and the image as input to train a neural network which, in turn, classifies whether a sentence was written by a human or a machine. Despite presenting good correlations with human judgments, learned metrics suffer from overfitting to particular domains and lack interpretability.

An important issue is the difficulty of evaluating the captions without enough reference captions to cover the diversity of vocabulary and visual content. Such a problem inspired the development of free-reference metrics, including UMIC, CLIPScore, and FAIEr. In such metrics, the visual content of images may be used to detect concepts, such as objects and the relationship between them, or compute a similarity measure between sentences and images using text-image pair network models.

Similar to the traditional metrics, novel reference-based metrics proposed for the image captioning task can be easily used to evaluate

² <https://github.com/tylin/coco-caption>

Table 2

Summary of evaluation metrics. Acronyms TR, FR, NG, WE, GR, VC, NN indicates, respectively, reference-based methods, free-reference methods, metrics based on n-gram comparison, word embedding-based metrics, metrics that model sentences in a semantic graph, methods that use visual content, and metrics trained using a neural network. Also, TK indicates the task: (I)Image Captioning, (V)ideo Captioning, (O)ther task.

N	Metric	TR	FR	NG	WE	GR	MM	LN	TK
1	BLEU (Papineni et al., 2002)	X		X					O
2	METEOR (Banerjee & Lavie, 2005)	X		X					O
3	CIDEr (Vedantam et al., 2015)	X		X					I
4	ROUGE (Lin, 2004)	X		X					O
5	SPICE (Anderson et al., 2016)	X				X			I
6	WMD (Kilickaya et al., 2017)	X			X				O
7	WEmbSim (Sharif, White et al., 2020)	X			X				I
8	VIFIDEL (Madhyastha et al., 2019)	X			X				I
9	LEIC (Cui et al., 2018)	X			X		X	X	I
10	NNEval (Sharif et al., 2018)	X			X			X	I
11	LCEval (Sharif et al., 2019)	X			X			X	I
12	TIGER (Jiang et al., 2020)	X			X		X		I
13	REO (Jiang et al., 2019)	X			X		X		I
14	BERTScore (Zhang et al., 2020)	X			X				I/O
15	ViLBERTScore (Lee et al., 2020)	X			X				I
16	SMURF (Feinglass & Yang, 2021)	X			X				I
17	CLIPScore (Hessel et al., 2021)		X		X		X		I
18	EMScore (Shi et al., 2022)		X		X		X		V
19	FAIEr (Wang et al., 2021)		X			X	X	X	I
20	UMIC (Lee et al., 2021)		X		X		X	X	I

video captioning approaches as they use only textual information. In fact, video and image captioning are similar tasks, since both require “translating” the visual content into a description in natural language. However, the critical difference between them is that video captioning requires taking into account the temporal information (actions). That is why metrics that use the visual content of images to compute a score cannot be easily extended to the video captioning task.

Recently, a reference-free metric named Emscore was proposed specifically for the video captioning task. It uses a video–text retrieval model that was pre-trained on more than 400 million image–text pairs. It can measure videos’ consistency with images and, effectively identifies “hallucinations” in captions.

A summary of the metrics presented in Fig. 3 is shown with more details in Table 2, which compares their key points investigated in this study.

5. Empirical experiments

This Section presents four simple empirical experiments to support a comparative analysis of the main characteristics and shortcomings of some selected metrics. First, we randomly selected some videos from popular video captioning datasets (see Section 3). Then, two hypothetical candidate sentences were created for each video: (a) a semantically incorrect candidate sentence using words present in the reference sentences, and (b) a semantically correct candidate sentence with words not present in the reference sentences. The experiments are detailed below, and all the code and data for reproducing these experiments will be available in Github.³

5.1. Popular metrics for video captioning

This analysis aims to examine the limitations of popular metrics used to evaluate video captioning approaches. We selected three video clips from different popular datasets (MSVD, MSR-VTT, and ActivityNet Captions) with related reference sentences (see Fig. 4).

It can be noticed that BLEU, METEOR, and CIDEr assigned high scores, highlighted in bold, to incorrect candidate sentences in all video clips. ROUGE-L assigned a better score for the correct sentence in Fig. 4. A because it contains the longest common subsequence compared to the wrong one. However, as seen in Fig. 4B and C, the sentence

with the longest common subsequence does not always sufficiently represent the visual content. All of these word-matching-based metrics fall short in their evaluation of these videos. This limitation comes from the fact that they place more weight on word-matching comparisons than they do on the vast diversity of linguistic expressions. The performance evaluation of video captioning algorithms using datasets containing only one or a few reference sentences, such as ActivityNet Captions or the Charades datasets, may not be sufficiently evaluated by such metrics due to that limitation.

Despite being designed to consider semantic content, the SPICE metric assigned the same score for correct and wrong semantic sentences in video clips Fig. 4.A and Fig. 4.C, and a lower score in Fig. 4.B. As briefly presented in Section 4.1.5, SPICE assigns a score by computing the similarity between encoded candidate and reference sentences in a semantic graph representation based on objects, attributes, and relations using a dependency tree parser. Despite considering synonymous in object nodes, it could not adequately evaluate those candidate sentences. This indicates that SPICE fails to evaluate the semantics when words are not similar between candidates and reference sentences.







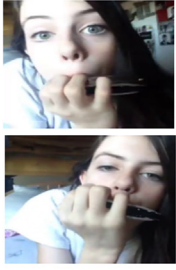


5.2. Potential metrics for video captioning

The purpose of this analysis is to look into the feasibility and accuracy of recently proposed image captioning metrics in the video captioning task. First, one video from ActivityNet Captions dataset was selected. Then, we created five more reference sentences, since there was only one sentence available for the video in the dataset. Some metrics compute the score using both the image and the reference sentences. For such metrics, we used the middle frame of the video, because it is more related to the reference sentences. Fig. 5 shows that candidate sentence A received a higher score (highlighted in bold) than candidate sentence B across all metrics, despite not adequately describing the image.

Also, BERTScore, WEmbSim, and SMURF all fail to assign a higher score to the correct sentence despite being proposed to consider semantic similarity using word embeddings.

Candidate A (correct sentence) received a higher score from TIGER, while both candidate sentences received a similar score from ViLBERTScore. To extract image–text pair features, these metrics employ a pre-trained neural network model. The final score is then computed using these features. Despite the fact that these models have been pre-trained on large datasets, they may be limited to the context in which they were trained.

³ <https://github.com/bioinfolab/survey-vidcap-metrics>

Scene	References	Candidate	B@1	B@2	B@3	B@4	R	S	M	C
A 	<ul style="list-style-type: none"> a cat chases a dog a cat is chasing a dog a cat is chasing around a german shepherd dog more than twice its size a cat is scaring away a dog a large dog approaches a small cat and a small cat is chasing a large dog in a garden the cat chased the dog around the cat is chasing the bigger dog away a dog and a cat are fighting a dog and cat is playing 	 a dog is chasing a cat	1.00	0.89	0.58	0.00	0.67	0.18	0.48	1.28
		 a cat is running after a dog	0.71	0.60	0.41	0.00	0.78	0.18	0.45	1.01
B 	<ul style="list-style-type: none"> a person walks with thick boots through the rough ways someone walks with heavy boots which are suitable for trucking there is someone advertising some boots a man talks about picking the right boot when you go hiking in the wilderness a man is walking in the forest area a pair of gray shows and a man walking through the snow with them guy walking with warm cloths and a bag in a snow area a person walks wearing a black trekking shore and climbed a snow mountain a mountain trekker emphasizes the importance of making a right choice about the trekking shoes you buy a man wearing heavy woolen cloths and boots walking on snow 	 a pair of gray hiking boots -is showing the forest area	0.91	0.67	0.53	0.37	0.54	0.12	0.28	0.61
		 a pair of hiking boots are shown and then a man walks along the snowy forest	0.76	0.44	0.29	0.00	0.52	0.09	0.27	0.22
C 	<ul style="list-style-type: none"> a girl is seen close up to a camera holding a harmonica closely to her mouth 	 a girl is seen close to a broken house	0.36	0.32	0.27	0.24	0.53	0.20	0.25	1.45
		 a woman seems to play the harmonica	0.12	0.00	0.00	0.00	0.24	0.20	0.07	0.44

 Correct candidate sentence  Wrong candidate sentence

Fig. 4. Example of popular metrics used to evaluate videos from different datasets. The video scenes are from (A) MSVD (video gjVBEJGHRk_26_38), (B) MSR-VTT (video video730), and (C) ActivityNet Captions dataset (video v_t1-GV2bAL4I). The first 10 sentence references from the original datasets were considered. In the columns, B@N, R, S, M and C denotes BLEU with N-grams (N = 1, 2, 3 and 4), ROUGE-L, SPICE, METEOR, and CIDEr-D, respectively. For further information about video caption datasets, see Section 3.

CLIPScore also assigned a higher score to the correct sentence. It is a free-reference metric that computes the score solely based on the visual content. This could imply that metrics that consider visual content evaluate semantics more effectively.

5.3. Specific metrics for video captioning

The purpose of this analysis is to assess the EMScore metric in particular. To the best of our knowledge, it is the only metric found for evaluating video captioning approaches to date. It compares the similarity of a video and a potential text as input. Fig. 6 illustrates a video from the MSR-VTT dataset selected for this experiment. Then, we calculated the similarity measure between nine sentences. Six of them were derived from the reference sentences of the original dataset. Other

three semantically incorrect sentences were created for the experiment. Notice that the wrong sentences (highlighted by a red background) achieved similar results to the correct ones (highlighted by a green background). Moreover, the lowest scored sentence “this is a video of a potato and a man” contains only the main concepts presented in the video (man and potato), but does not consider the action performed. This fact indicates that the missing information (action or objects) influences the metric’s score.

Considering that videos may contain audio information and that such data may be essential to describe a given video adequately, we selected another video from the MSR-VTT dataset to analyze such a scenario, as shown in Fig. 7. Six sentences were extracted from the reference sentences of the original dataset (highlighted by a green background). The other three semantically wrong sentences were created (highlighted by a red background).



- References**
- The person cleans a gym with the large dust mop
 - A man sweeps the floor of a gym
 - A man sweeps the gymnasium floor with a dust broom
 - A man in a blue shirt and black trousers sweeping a gym floor
 - A man in a blue shirt is sweeping a classroom
 - A man in black pants is mopping the floor of a gym room

Candidate A
A man in a blue shirt and black pants is working out in a gym

Reference based	Word matching	BLEU_4	0.60
		METEOR	0.44
		ROUGE_L	0.74
		CIDEr_D	1.25
	Scene graph	SPICE	0.50
	Word embedding	BERTScore	0.75
		SMURF	0.98
		WEEmbSim	0.29
	Visual Content	VILBERTScore	0.90
		TIGEr	0.75
Reference free	Hand crafted	CLIPScore	0.69

Candidate B
A person is mopping the gym floor

Reference based	Word matching	BLEU_4	0.00
		METEOR	0.26
		ROUGE_L	0.47
		CIDEr_D	0.79
	Scene graph	SPICE	0.15
	Word embedding	BERTScore	0.70
		SMURF	0.65
		WEEmbSim	0.22
	Visual Content	VILBERTScore	0.89
		TIGEr	0.78
Reference free	Hand crafted	CLIPScore	0.74

Fig. 5. Analysis of scores given by evaluation metrics with two candidate sentences. Candidate A is a semantically incorrect sentence, although it contains words present in the reference sentences. Candidate B is a semantically correct sentence, but it does not contain words in the same order as those presented in the reference sentences.

Video	Sentences	EMScore
	a guy slices a potato	0,31
	a man is cutting a potato	0,31
	a man slices a potato in the kitchen	0,32
	a man is cutting a potato in a kitchen with knife and talking about that	0,32
	a man shows how to thinly slice potatoes	0,35
	a man slices potato into equal pieces	0,32
	potatoes are slicing the man	0,32
	a man demonstrates how to fry thin potatoes slices	0,32
	this is a video of a potato and a man	0,29

Fig. 6. Example of evaluation scores assigned by the EMScore metric for a video presented in MSR-VTT dataset. The reference sentences presented in the dataset are in green, whilst, those in red are semantically incorrect candidate sentences. The best score is highlighted in bold. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Observe that EMScore assigns the highest score to a semantically incorrect sentence. Sentences that describe only parts of the visual content achieved similar scores (between 0.27 and 0.29). The sentence with the lowest score describes a man showing the food while another man (who is not visualized in the scene) provides comments. Such behavior highlights that the EMScore metric does not consider audio information and, thus, could not fully match the description with visual information.

Video	Sentences	EMScore
	video of a man talking about an overweight man	0,30
	an obese man is wearing a red sweater	0,29
	a man is talking about how much a man weighs and what he has done	0,27
	a male showing off his food while another male provides commentary	0,22
	a guy talks about how another man is so fat	0,28
	a man holds a big plate of food and he weighs 910 pounds	0,28
	this is a video of a man watching TV	0,26
	a fat man wearing a black shirt is cooking a sandwich in front of a camera	0,28
	a video showing an obese man wearing a red sweater speaking in a documentary	0,32

Fig. 7. Example of the evaluation scores assigned by EMScore for a video found in the MSR-VTT dataset which considers the audio. Sentences in green are reference sentences presented in the dataset. The sentences in red are candidate semantically incorrect sentences. The best score is highlighted in bold. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5.4. Analysis

Based on the three experimental evaluations shown before, we observed that most metrics still fail to assess semantic aspects from visual descriptions. Also, the metrics based on word embeddings still have space for improvements, considering that they are based on the corpus on which they were trained. Many word embedding models only generate a fixed feature vector for each word. However, a word can have a different semantic meaning depending on where it occurs in a

sentence. In addition, missing words in word embedding models may result in a low score.

When trained on a biased dataset, metrics that use pre-trained neural network models to extract visual and textual features may suffer from performance degradation. An example of this is shown in Fig. 6, where the sentence that begins with “A man demonstrates how” achieved a high score. Apparently, this happened since the sentence had a sequence of words with a high frequency in the pre-trained dataset. Accordingly, Caglayan, Madhyastha, and Specia (2020) also reported that some metrics tend to produce unexpected high scores in some benchmarks when using the most frequent sentence in the training set.

Furthermore, during the experiments, we also observed that there are several datasets available for evaluating videos and image captioning tasks with different aspects (Aafaq et al., 2019). Some datasets have videos with temporal discontinuities, such as sudden changes in action or appearance, which may affect negatively the score of reference-free metrics, as shown in Fig. 6.

We also noticed that datasets, such as MSR-VTT and ActivityNetCaptions, make available videos with audio. In these cases, the reference sentences may consider both audio and video to describe the video scene. For example, in Fig. 4(B), the reference sentences “There is someone advertising some boots” and “A pair of gray hiking boots are shown as the narrator states boots are the most important part of hiking” were created taken into account both, the visual and audio information. Similar candidate sentences would be scored low by EMScore as they only use the visual content for the evaluation.

Finally, the number of reference sentences available in the datasets seems to influence the computation of some metrics, as reported in previous works (Jiang et al., 2020; Madhyastha et al., 2019; Sharif, Nadeem et al., 2020). Those findings may indicate that some metrics may not achieve good results when used in specific datasets, especially when there are few reference sentences associated to each video.

6. Discussion

6.1. Limitations of the evaluation metrics

Traditional metrics, such as BLEU, METEOR, and ROUGE-L, are commonly used to evaluate the performance of image and video captioning approaches by means of a simple and quick procedure. However, the main weakness of those metrics is that they are based on n-gram overlapping, which compares a candidate sentence with human-written reference sentences. Therefore, they are highly dependent on how the words appear in the reference sentences for evaluating a candidate sentence, which must be generated in the same order and using the same words presented in the reference sentences to achieve high scores.

CIDEr was the first metric proposed specifically for evaluating image captioning approaches. It introduced a novel paradigm that propose to measure the consensus of the human judgment. Although CIDEr improved the accuracy over the existing metrics, it is also based on n-grams comparison and suffers from the same problems reported before. Besides, it also does not take into account the semantic information contained in the sentences.

Due to the high dependency on correct n-gram matching and the difficulty of assessing the semantics of sentences and words by the above-mentioned metrics, SPICE was devised to evaluate the semantic content of descriptions automatically generated for images. Actually, it was able to satisfactorily measure the semantics between a candidate sentence and reference sentences by creating graph-based semantic representations, which are not considered by other metrics. However, this metric is highly dependent on a semantic parser, since it fails to perform the lexical and syntactical evaluation of the generated sentences.

Furthermore, SPICE computes the semantics by measuring string matching, making the evaluation challenging to scale or adapt to different languages and domains (Madhyastha et al., 2019). Also, unpredictable failures can be caused by sentence parsing issues or problems in created semantic representations (Feinglass & Yang, 2021).

In the last few years, many metrics have been appeared for evaluating image and video captioning systems, depicted in Fig. 3 and detailed in Table 2.

Motivated by the WMD metric, some metrics, including WEembSim, VIFIDEL, and BERTScore were proposed to tackle the problem of evaluating the semantic meaning between words or sentences using word embeddings. However, although word embeddings can provide some semantic representations of words, they may introduce biases to the evaluation process once they are learned using a specific corpus.

The visual content of images was also considered in some metrics (Cui et al., 2018; Hessel et al., 2021; Jiang et al., 2019, 2020; Lee et al., 2021, 2020) by encoding the visual and text data into a common semantic vector space using a pre-trained model. The SCAN (Stacked Cross Attention Neural Network) (Lee et al., 2018), a network pre-trained on the 2014 MS-Coco dataset proposed for the image-text matching problem, is commonly used (Jiang et al., 2019, 2020).

Moreover, some approaches consider the visual content during evaluation by detecting objects using a pre-trained object detector model, for instance, the Faster R-CNN model (Madhyastha et al., 2019; Wang et al., 2021). Despite achieving high correlation with the human judgment, they are also highly dependent on pre-trained models. Besides, some of them may overlook the syntactic correctness of the captions and their relevance to the image. Thus, it is advisable to combine them with other metrics (Stefanini et al., 2023).

Learned metrics have emerged more recently (Cui et al., 2018; Sharif et al., 2019, 2018; Wang et al., 2021). They use neural networks mainly trained to distinguish between human and machine-generated captions. The concern about these metrics is that they are “gameable” that is, susceptible to manipulation. This means they can be used as an objective function for training video captioning approaches, achieving high scores while still generating syntactically and/or semantically incorrect sentences (Gao, Galley, & Li, 2019).

Although most of these metrics are robust and present a good correlation with human judgments, they output only a single score to assess the quality of the captions generated by the system. However, a single value may not provide enough information to interpret the low quality of a given system or to explain specific errors. In other words, metrics, in general, lack ways to provide human-comprehensible explanations of their meaning.

REO is the first metric that tackles the interpretability problem by computing a score involving three aspects: relevance regarding ground truth, extra description beyond image content, and omitted ground truth information. Despite providing a score for each aspect, it is a visual-content metric that uses a pre-trained model to extract feature vectors, but does not present a clear explanations of the scores.

Inspired by the drawbacks presented in metrics used for video captioning evaluation, a free-reference metric called EMSCore was proposed to measure the similarity between a video and a candidate sentence. It is an embedding-based evaluation metric and uses the pre-trained image-language model CLIP (Radford et al., 2021) to obtain image and text embeddings. Coarse-grained (video level) and fine-grained (frame level) embeddings are combined to get characteristics of the visual elements of the video over time. Despite being promising, such metric consider only the visual content of videos and may fail to evaluate approaches trained with multimodal data (audio and visual information) currently available in some recent datasets. For instance, Fig. 4B) presents some reference sentences considering the narrator’s speech. A video captioning system trained on such a dataset with multimodal data will generate similar sentences that would probably be scored lower while using the EMSCore metric. Although EMSCore does not need reference sentences, the authors recommend using them when available, as they are complementary and may lead to information gain.

Despite the large number of metrics reported in the literature, none of them has been widely adopted. Many of these metrics are restricted to evaluating image description systems and cannot be satisfactorily extended to the video captioning task. Moreover, the use of pre-trained models to extract features that were previously trained in specific contexts may fail to represent a video from another context. Even if a generic pre-trained model is used, it may fail as the input may have unknown situations or vocabulary.

6.2. Possible extension to other languages

Most evaluation metrics use English language resources, such as dependency tree parsers, synonymous dictionaries, or pre-trained neural networks to capture semantic or syntax information from the sentences. Such metrics depend on such an apparatus, making extending them to other languages difficult or unfeasible. However, due to their simplicity, n-gram-matching-based metrics (BLEU, METEOR, CIDEr, and ROUGE) have been used for languages other than English, such as Italian (Antonio, Croce, & Basili, 2019), Hindi (Singh, Singh, & Bandyopadhyay, 2022), Portuguese (dos Santos, Colombini, & Avila, 2022), and Chinese (Liu, Hu, Li, Yu, & Guan, 2020).

Also, the metrics based on word embedding cannot be easily extended to other languages. They require pre-trained word embedding models in the target language, which, frequently, are not available. The same holds for reference-free metrics, which use features extracted from pre-trained network models for the visual-word matching task. Since there are not many publicly available datasets of this type for other languages, the use of such metrics is also limited to the English language only.

Scene graph-based metrics such as SPICE and FAIEr require a language parser to detect the concepts and relationships between objects and subjects. This metric was successful due to the significant advances obtained with such tools in the English language, for example, the Stanford parser. However, extending these metrics to other languages is challenging due to the difficulty of finding similar tools in other languages.

7. Conclusions and research trends

Image and video caption evaluation is a complex task that involves semantics and matching of the visual content and text. In the recent years, many evaluation metrics were proposed, aiming at circumventing the drawbacks and challenges faced by their preceding approaches.

In the present study, a survey on automatic evaluation metrics for video captioning was done. We proposed a taxonomy, categorizing the metrics and discussing their pros and cons. Additionally, this study also analyzed the existing metrics, pointing out their main weaknesses.

It was noticed that most of those metrics, presented in Section 4, were proposed to address specific shortcomings of previous metrics, including the lack of semantic evaluation, insufficient reference sentences, poor correlation with human judgments, lack of generalizability, and lack of explainability. Notably, these metrics focus on achieving a strong correlation with human judgments while overlooking other desirable characteristics, including computational cost, bias, consistency, sensitivity, and ease of use. As such, further research is required to develop metrics that cover the desirable characteristics for evaluating video description systems.

We hope this research will provide a reference for researchers to understand the current drawbacks and advantages of the existing metrics for image and video captioning and new insights for developing new metrics.

Based on the deep analysis of the main drawbacks of the metrics, the advancement of the state-of-the-art in the field of image and video captioning evaluation will require extensive research efforts towards the following directions:

- **Semantics:** Existing metrics often fail to evaluate the semantics of the visual content since it can be described by many different sentences written in natural language. Evaluating semantic similarity among those sentences or between a sentence and visual content is challenging. Reference-based metrics usually use word-matching or word-embedding features to estimate semantic similarity and often neglect visual relevance and details. Although free-reference metrics have presented promising results in extracting features and inferring semantic matching, they often neglect the syntactic structure of sentences. Also, they may result in a biased evaluation since they use pre-trained models and are limited by the training data context. Further research is needed to find ways to evaluate semantics in the computational scenario. More specifically, the semantics of the visual content of videos and the semantics of texts that describe the video contents. In both cases, such metrics should ideally measure the semantics of the complex interactions between entities and objects over time.
- **Explainability:** Existing metrics used to evaluate video captioning approaches only provide a single score. Recent state-of-the-art comparisons report these scores sometimes with significant differences. Usually, no explanation are provided for such a behavior. In fact, a single score cannot provide a meaningful interpretation or intuition about why and when an approach is better than another. Furthermore, although efficient for some tasks, neural network-based approaches are “black-boxes” that lack traceability and transparency of their computed results. Therefore, an ideal metric should provide, in addition to an overall system score, information about errors made by the system (e.g., hallucinations, missing information, incorrect subject/action/object ratio). In this context, comprehensive human-comprehensible metrics require explainability, which should ideally comply with the principles of explainable artificial intelligence proposed by Phillips et al. (2021). Certainly, this is a challenging endeavor that will require interdisciplinary research.
- **Adaptability:** While generic captioning systems have been evaluated using current measures, context-specific techniques can call for additional measurements. When captioning medical photos, for instance, the generated captions should ideally help with the diagnosis, and the resulting medical report should not include descriptions of the image’s components that are not relevant for the diagnosis. As a result, programs that are specialized to a certain context should adopt appropriate evaluation metrics.
- **Extension to other languages:** Some metrics, especially those that aim to capture the semantic aspects of the video, use features extracted from neural networks trained on a specific corpus or specific language parsers. However, they cannot be easily extended to languages other than English, as discussed in Section 6.2. Future research may include creating such language resources, allowing the extension of some metrics to languages other than English. However, due to the wide differences in word and expression meanings and grammatical differences across modern languages, we do not foresee a language-agnostic reference-free metric emerging soon.
- **Datasets:** When there are few reference sentences available, some measures have a poor performance. Therefore, high-quality datasets with multiple reference sentences are essential to improve evaluation reliability. The MSVD and MSR-VTT datasets, which feature numerous annotations per video, are the most frequently employed in the video description task. However, many of those videos have points of discontinuity in the scenes that can negatively affect the performance of reference-free metrics, such as motion or scene switching. The ideal reference scenario would be a complete (for a given domain) and high-quality “gold standard” dataset with many reference sentences that adequately describe the video scene differently. A dataset like this one could facilitate the creation of new reference-based and reference-free measures and establish a standard by offering a precise, consensual assessment of the effectiveness of video captioning.

- **Multimodal free-reference metrics:** Since a video combines both audio and visual information, the audio may be necessary to effectively communicate the video's content. To the best of our knowledge, the EMScore is the only metric proposed for evaluating video captioning approaches and computes a similarity score between a video (visual information captured from frames) and a sentence. However, a video contains, beyond visual information, audio information, which may be essential to describe a video scene. For instance, consider a video scene of a woman sitting in a chair giving an interview about education issues, and the following candidate sentences: (a) "A woman is sitting in a chair and moving her hands", and (b) "A woman is sitting in a chair and talking about education issues". The EMScore provides a higher score to the first candidate sentence, even though the second candidate sentence better describes the given video. Previous studies (Hori et al., 2017; Ramanishka et al., 2016) have shown that combining audio features, such as MFCC, and visual features can improve the performance of video captioning approaches. Thus, a potential future work should investigate novel reference-free metrics capable of including, in addition to visual information, audio information (when available) in the evaluation of video descriptions.

CRedit authorship contribution statement

Andrei de Souza Inácio: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. **Heitor Silvério Lopes:** Conceptualization, Writing – original draft, Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

A.S. Inácio thanks UNIEDU/FUMDES – Pós-Graduação for the scholarship, and H.S. Lopes thanks the Brazilian National Research Council (CNPq) for the research grant 311785/2019-0.

References

Aafaq, N., Mian, A., Liu, W., Gilani, S. Z., & Shah, M. (2019). Video description: A survey of methods, datasets, and evaluation metrics. *ACM Computing Surveys*, 52(6), 1–37.

Amirian, S., Rasheed, K., Taha, T. R., & Arabnia, H. R. (2020). Automatic image and video caption generation with deep learning: A concise review and algorithmic overlap. *IEEE Access*, 8(1), 218386–218400.

Anderson, P., Fernando, B., Johnson, M., & Gould, S. (2016). Spice: Semantic propositional image caption evaluation. In *European conference on computer vision* (pp. 382–398).

Antonio, S., Croce, D., & Basili, R. (2019). Large scale datasets for image and video captioning in Italian. *Italian Journal of Computational Linguistics*, 5(5–2), 49–60.

Baãzaoui, A., Barhoumi, W., Ahmed, A., & Zagrouba, E. (2018). Modeling clinician medical-knowledge in terms of med-level features for semantic content-based mammogram retrieval. *Expert Systems with Applications*, 94, 11–20.

Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proc. of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization* (pp. 65–72).

Bin, Y., Shang, X., Peng, B., Ding, Y., & Chua, T.-S. (2021). Multi-perspective video captioning. In *Proc. of the 29th ACM international conference on multimedia* (pp. 5110–5118).

Caglayan, O., Madhyastha, P. S., & Specia, L. (2020). Curious case of language generation evaluation metrics: A cautionary tale. In *Proc. of the 28th international conference on computational linguistics* (pp. 2322–2328).

Chandrasekaran, D., & Mago, V. (2021). Evolution of semantic similarity – a survey. *ACM Computing Surveys*, 54(2), 1–37.

Chen, D., & Dolan, W. (2011). Collecting highly parallel data for paraphrase evaluation. In *Proc. of the 49th annual meeting of the association for computational linguistics* (pp. 190–200).

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., et al. (2015). Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325.

Chen, Y.-C., Li, L., Yu, L., El Kholi, A., Ahmed, F., Gan, Z., et al. (2020). Uniter: Universal image-text representation learning. In *European conference on computer vision* (pp. 104–120).

Cui, Y., Yang, G., Veit, A., Huang, X., & Belongie, S. (2018). Learning to evaluate image captioning. In *Proc. of IEEE conference on computer vision and pattern recognition* (pp. 5804–5812).

Denkowski, M., & Lavie, A. (2010). Choosing the right evaluation for machine translation: An examination of annotator and automatic metric performance on human judgment tasks. In *Proc. of the 9th conference of the association for machine translation* (pp. 1–9).

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of the conference of the north american chapter of the association for computational linguistics* (pp. 4171–4186).

dos Santos, G. O., Colombini, E. L., & Avila, S. (2022). #PraCegoVer: A large dataset for image captioning in Portuguese. *Data*, 7(2), 1–27.

Feinglass, J., & Yang, Y. (2021). SMURF: SeMantic and linguistic UndeRstanding fusion for caption evaluation via typicality analysis. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)* (pp. 2250–2260).

Gao, J., Galley, M., & Li, L. (2019). Neural approaches to conversational AI. *Foundations and Trends in Information Retrieval*, 13(2–3), 127–298.

Heilbron, F. C., Escorcia, V., Ghanem, B., & Niebles, J. C. (2015). ActivityNet: A large-scale video benchmark for human activity understanding. In *Proc. of IEEE conference on computer vision and pattern recognition* (pp. 961–970).

Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., & Choi, Y. (2021). CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 7514–7528).

Hori, C., Hori, T., Lee, T.-Y., Zhang, Z., Harsham, B., Hershey, J. R., et al. (2017). Attention-based multimodal fusion for video description. In *Proc. of the IEEE international conference on computer vision* (pp. 4193–4202).

Inácio, A. D. S., Gutoski, M., Lazzaretti, A. E., & Lopes, H. S. (2021). OSVidCap: A framework for the simultaneous recognition and description of concurrent actions in videos in an open-set scenario. *IEEE Access*, 9, 137029–137041.

Jain, V., Al-Turjman, F., Chaudhary, G., Nayar, D., Gupta, V., & Kumar, A. (2022). Video captioning: A review of theory, techniques and practices. *Multimedia Tools and Applications*, 81(25), 35619–35653.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4), 422–446.

Ji, W., & Wang, R. (2021). A multi-instance multi-label dual learning approach for video captioning. *ACM Transactions on Multimedia Computing, Communications and Applications*, 17(2s), 1–18.

Jiang, M., Hu, J., Huang, Q., Zhang, L., Diesner, J., & Gao, J. (2019). REO-relevance, extraness, omission: A fine-grained evaluation for image captioning. In *Proc. of 10th international joint conference on natural language processing* (pp. 1475–1480).

Jiang, M., Huang, Q., Zhang, L., Wang, X., Zhang, P., Gan, Z., et al. (2020). Tiger: Text-to-image grounding for image caption evaluation. In *Proc. 9th international joint conference on natural language processing* (pp. 2141–2152).

Kilickaya, M., Erdem, A., Ikizler-Cinbis, N., & Erdem, E. (2017). Re-evaluating automatic metrics for image captioning. In *Proc. of the 15th conference of the european chapter of the association for computational linguistics* (pp. 199–209).

Krishna, R., Hata, K., Ren, F., Fei-Fei, L., & Niebles, J. C. (2017). Dense-captioning events in videos. In *Proc. of the IEEE international conference on computer vision* (pp. 706–715).

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1), 79–86.

Laina, I., Rupprecht, C., & Navab, N. (2019). Towards unsupervised image captioning with shared multimodal embeddings. In *Proc. of the IEEE/CVF international conference on computer vision* (pp. 7414–7424).

Lee, K.-H., Chen, X., Hua, G., Hu, H., & He, X. (2018). Stacked cross attention for image-text matching. In *Proc. of the european conference on computer vision* (pp. 212–228).

Lee, H., Yoon, S., Deroncourt, F., Bui, T., & Jung, K. (2021). UMIC: An unreferenced metric for image captioning via contrastive learning. In *Proc. of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 2: short papers)* (pp. 220–226).

Lee, H., Yoon, S., Deroncourt, F., Kim, D. S., Bui, T., & Jung, K. (2020). ViLBERTScore: Evaluating image caption using vision-and-language BERT. In *Proc. of the 1st workshop on evaluation and comparison of NLP systems* (pp. 34–39).

- Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74–81).
- Liu, M., Hu, H., Li, L., Yu, Y., & Guan, W. (2020). Chinese image caption generation via visual attention and topic modeling. *IEEE Transactions on Cybernetics*, 52(2), 1247–1257.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. ArXiv, arXiv:1907.11692.
- Liu, X., Xu, Q., & Wang, N. (2019). A survey on deep neural network-based image captioning. *The Visual Computer*, 35(3), 445–470.
- Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in neural information processing systems* (pp. 1–11).
- Madhyastha, P. S., Wang, J., & Specia, L. (2019). VIFIDEL: Evaluating the visual fidelity of image descriptions. In *Proc. of the 57th annual meeting of the association for computational linguistics* (pp. 6539–6550).
- Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proc. of the 40th annual meeting of the association for computational linguistics* (pp. 311–318).
- Perlin, H. A., & Lopes, H. S. (2015). Extracting human attributes using a convolutional neural network approach. *Pattern Recognition Letters*, 68, 250–259.
- Phillips, P. J., Hahn, C. A., Fontana, P. C., Yates, A. N., Greene, K., Broniatowski, D. A., et al. (2021). *Four principles of explainable artificial intelligence: Internal report NISTIR 8312*, National Institute of Standards and Technology.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al. (2021). Learning transferable visual models from natural language supervision. In *Proc. of the 38th international conference on machine learning*, Vol. 139 (pp. 8748–8763).
- Rafiq, M., Rafiq, G., & Choi, G. S. (2021). Video description: Datasets & evaluation metrics. *IEEE Access*, 9, 121665–121685.
- Ramanishka, V., Das, A., Park, D. H., Venugopalan, S., Hendricks, L. A., Rohrbach, M., et al. (2016). Multimodal video description. In *Proc. of the 24th ACM international conference on multimedia* (pp. 1092–1096).
- Rohrbach, A., Rohrbach, M., Qiu, W., Friedrich, A., Pinkal, M., & Schiele, B. (2014). Coherent multi-sentence video description with variable level of detail. In *Proc. of the 36th german conference on pattern recognition* (pp. 184–195).
- Rohrbach, A., Torabi, A., Rohrbach, M., Tandon, N., Pal, C., Larochelle, H., et al. (2017). Movie description. *International Journal of Computer Vision*, 123, 94–120.
- Sharif, N., Nadeem, U., Shah, S. A. A., Bennamoun, M., & Liu, W. (2020). Vision to language: Methods, metrics and datasets. In *Machine learning paradigms* (pp. 9–62).
- Sharif, N., White, L., Bennamoun, M., Liu, W., & Shah, S. A. A. (2019). Lceval: Learned composite metric for caption evaluation. *International Journal of Computer Vision*, 127(10), 1586–1610.
- Sharif, N., White, L., Bennamoun, M., Liu, W., & Shah, S. A. A. (2020). WEmbSim: A simple yet effective metric for image captioning. In *Proc. of IEEE digital image computing: techniques and applications* (pp. 1–8).
- Sharif, N., White, L., Bennamoun, M., & Shah, S. A. A. (2018). NNEval: Neural network based evaluation metric for image captioning. In *Proc. of the european conference on computer vision* (pp. 37–53).
- Shi, Y., Yang, X., Xu, H., Yuan, C., Li, B., Hu, W., et al. (2022). EMScore: Evaluating video captioning via coarse-grained and fine-grained embedding matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 17929–17938).
- Sigurdsson, G. A., Varol, G., Wang, X., Farhadi, A., Laptev, I., & Gupta, A. (2016). Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Proc. of the european conference on computer vision* (pp. 510–526).
- Singh, A., Singh, T. D., & Bandyopadhyay, S. (2022). Attention based video captioning framework for Hindi. *Multimedia Systems*, 28(1), 195–207.
- Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., & Cucchiara, R. (2023). From show to tell: A survey on deep learning-based image captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1), 539–559. <http://dx.doi.org/10.1109/TPAMI.2022.3148210>.
- Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). Cider: Consensus-based image description evaluation. In *Proc. of the IEEE conference on computer vision and pattern recognition* (pp. 4566–4575).
- Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., & Saenko, K. (2015). Sequence to sequence – video to text. In *Proc. of the IEEE international conference on computer vision* (pp. 4534–4542).
- Wang, S., Yao, Z., Wang, R., Wu, Z., & Chen, X. (2021). FAIEr: Fidelity and adequacy ensured image caption evaluation. In *Proc. of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14050–14059).
- Witten, I. H., & Frank, E. (2005). *Data mining: practical machine learning tools and techniques* (2nd ed.). Morgan Kaufmann.
- Xu, J., Mei, T., Yao, T., & Rui, Y. (2016). MSR-VTT: A large video description dataset for bridging video and language. In *Proc. of the IEEE conference on computer vision and pattern recognition* (pp. 5288–5296).
- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. In *Proc. 8th international conference on learning representations* (pp. 1–43).
- Zhou, L., Xu, C., & Corso, J. J. (2018). Towards automatic learning of procedures from web instructional videos. In *Proc. of the 32nd AAAI conference on artificial intelligence* (pp. 7590–7598).



Andrei de Souza Inácio received the BSc and MSc degree in Computer Science from the Federal University of Santa Catarina (UFSC) in 2013 and 2016, respectively. Since 2014, he has been a lecturer at the Federal Institute of Santa Catarina (IFSC). He is currently pursuing a Ph.D. degree in Electrical and Computer Engineering at the Federal University of Technology – Paraná, PR, Brazil. He has professional experience in information systems design, web development, and IT project management. His research interests include, but are not limited to, computer vision, machine learning, and data mining.



Heitor Silvério Lopes received the BSc and MSc degrees in Electronic Engineering from Federal University of Technology – Paraná (UTFPR) in 1984 and 1990, respectively, and his PhD from the Federal University of Santa Catarina in 1996. Later, in 2014, he spent a sabbatical year at the Department of Electrical Engineering and Computer Science at the University of Tennessee, USA. Since 2003, he has been a research fellow of the Brazilian National Research Council in the area of Computer Science. Currently, he is a tenured full Professor with the Department of Electronics and the Graduate Program in Electrical Engineering and Applied Computer Science (CPGEI) at UTFPR, Curitiba. He was the co-founder and former president of the Brazilian Society for Computational Intelligence (SBIC). His major research interests are in the fields of computer vision, deep learning, evolutionary computation, and data mining.