EVALUATION OF AUTOMATICALLY GENERATED VIDEO CAPTIONS USING VISION AND LANGUAGE MODELS

Luis Lebron⋆ *Yvette Graham*† *Noel E. O'Connor*⋆ *Kevin McGuinness*⋆

Insight SFI Research Centre for Data Analytics, Dublin City University (DCU)⋆;
School of Computer Science and Statistics, Trinity College Dublin†
luis.lebroncasas@insight-centre.org

ABSTRACT

Vision and language models are easily transferred to other tasks. In particular, they have been demonstrated to work well in the evaluation of automatic image captioning. This has made it possible to evaluate systems without the need for references or additional information apart from the image and the caption. However, these models do not provide a straightforward way of evaluating videos. In this paper, we propose using these models for video captioning evaluation. We explore the use of both single image-based evaluation and different methods to include data from multiple frames. Experiments demonstrate that using clustering methods to select a few frames to compute the final score gives an excellent correlation with human judgment. The bias in the human annotations can also influence the metric, so we propose filtering the human assessments to discard outliers and improve the evaluation process.

***Index Terms*—** Video Captioning Evaluation, Vision and Language Models, Clustering, Transformers, Video Captioning

1. INTRODUCTION

Traditionally video captioning and image captioning are evaluated by comparing the captions against references written by humans. Some examples of these metrics are BLEU [1], ROUGE [2], METEOR [3], or CIDEr [4]. All of these metrics are based on the computation of precision or recall using n-grams between the candidate caption and the set of references. Some approaches improve upon this idea using stemming or synonyms e.g. SPICE [5] uses a scene graph built from the references to evaluate the caption. However, all approaches use human annotations to perform the evaluation.Furthermore, all approaches fail to account for the visual information in the image.

The next evolution of approaches are learned metrics where a model is trained to evaluate the caption. For example, BERTScore [6] is based on the popular BERT [7], retrained to evaluate machine translation sentences. Similarly, BERTHA [8] is also a language model trained to evaluate video captioning using only text data. An improvement to this idea is to include the visual information of the image. TIGEr[9] use pretrained models to evaluate an image caption using both textual and visual information. LEIC[10] uses a CNN and an LSTM to extract the visual and textual information and combines them to detect if the caption is written by a human, removing the need for human references.

Recent developments have explored combining image and text in a shared space in more depth. These models are trained in very large image and caption pairs datasets using contrastive learning to find image embeddings directly from the associated language. The objective of these models, like CLIP [11] and CLOOB [12], is commonly focused on the task of zero-shot learning. This goal makes it easy to transfer the model without retraining.

Taking this idea, CLIPScore [11] proposes various setups to evaluate image captioning using CLIP [13], a well-known vision and language model. They propose using similarity between embeddings of the caption and the image to rank the scores. They obtain state of the art results in multiple image captioning datasets. Similarly, UNITER [14] also proposes a vision and language model pretrained on pairs of captions and images, obtaining similar results to CLIPScore. These approachs prove to better correlate with human judgment than previous metrics and have the advantage that they do not need human references. Both of them perform well in image captioning evaluation; however, they are limited to only using a single image.

Our goal is to transition these ideas into video captioning evaluation. Video captioning is a similar task as image captioning evaluation; however, more factors need to be considered as the input contains more information than a single image. We first test two vision and language models against image and video captioning evaluation datasets to study their different performances in a simple scenario of a single image and caption. The experiments reported here then explore different strategies to use the frames, from using a single frame as input to computing the mean of the embedded frames to a more complex experiment where the frames are clustered.

The finetuning of these models is also studied to further adapt them to the task. We obtain better results than the other metrics in various setups with the best results achieved using clustering and model retraining. Finally, the outliers and the human bias in the annotations could affect the quality of the evaluation, more noticably when training a new model. The final set of experiments tries to filter the human judgments to find the annotations that do not follow the same pattern as the others. We discover that removing only a small set of captions improves all the metrics tested.

2. PROPOSED METHOD

2.1. Vision and Language Models

The two cross modal vision and language embedding models we use are CLIP [13] and CLOOB [12]. They follow a similar structure: both of them have an image encoder and text encoder that are trained in huge datasets of image-caption pairs extracted from the web. Both use a Transformer as a text encoder; the difference is that the recommended CLIP uses a Vision Transformer as the image encoder, while CLOOB is built based on ResNet image encoder.

CLIP is trained to maximise the cosine similarity of the two vectors while minimising the similarity of the mismatch pairs using InfoNCE [13] CLOOB instead uses InfoLOOB [12], which aims to remove the explaining away problem of InfoNCE.

2.2. Evaluation methodology with a single image

Following the same methodology as [11] we use the image and text encoder of the vision and language model and compute the cosine similarity between the two vectors.

$$s(c,t) = \cos(\mathrm{VLM}_{image}(c), \mathrm{VLM}_{text}(t)), \tag{1}$$

where c is the image, t is the textual caption and VLM is the vision and language model divided in the image and text encoder, and

$$\cos(u,v) = \frac{u^T v}{\|u\|\|v\|}$$

is the cosine similarity between embeddings u and v. In the most simple scenario, the setup is the same for image captioning and video captioning. In the case of video, only the middle frame is used. An additional advantage of this metric compared to traditional metrics is that it does not require human references to evaluate the caption.

This configuration can be extended to use multiple references. In this case the caption is compared to the references, unlike in [11] where they compare the references to the image. Then the maximum similarity is added to the previous similarity score between the image and the caption:

$$\begin{aligned} s(c,t,R) = &\cos(\mathrm{VLM}_{image}(c), \mathrm{VLM}_{text}(t)) \\ &+ \max_{r\in R} \left\{\cos(\mathrm{VLM}_{text}(t), \mathrm{VLM}_{text}(r))\right\}, \end{aligned} \tag{2}$$

where R is the set of references.

2.3. Evaluation methodology for a video

In the cases of a video, the visual information comes from multiple frames; however, none of the vision and language models have a mechanism to include all the frames simultaneously. The first idea is to compute the similarity of the caption for all the frames and take the maximum similarity. The same idea can be applied with the mean and minimum similarity:

$$s(F,t) = \mathrm{op}_{f\in F}\left\{\cos(\mathrm{VLM}_{image}(f), \mathrm{VLM}_{text}(t))\right\}, \tag{3}$$

where $\mathrm{op} \in \{\min, \max, \mathrm{mean}\}$ and F is the set of frames. Videos are usually composed of a few shots, so it is unnecessary to compute the similarity between all the frames in the video. Some of these frames may be noise or a blank frame between shots, producing undesirable results in terms of the similarity with the caption. To reduce the effect of these issues, we study clustering and classification techniques to detect similar frames. The first method we test is to find the k-NN [15] frames to the caption and compute the similarity with them. Using multiple frames mitigates the effect of possible noisy frames in the evaluation. It also only takes frames that are closely described by the caption and not any other random action that occurs in the video. The second method we test is based on clustering where we use k-means [16] to identify the different shots in the video and obtain a small set of frames that represent the main parts of the video.

2.4. Training models

One reasonable step to continue with this line of work is to explore if the pretrained model can be fine tuned to the specific video captioning evaluation task. We use the same loss functions as in each original paper. We keep most of the hyperparameters but change the learning rate and the weight decay to a lower one.

For the data, we took all the TRECVid-VTT dataset 2016 to 2020 [8] and reserved 2020 for testing, using the rest of the years as train/validation. During training, we found that using the predicted captions did not improve the final metric, which was discarded early, and only the references were used. In terms of the size of the dataset, only using four years of data for training does not provide enough pairs to train the model. To increase the amount of data, we augment the text. However data augmentation is not used at the image level, all the video frames are used as separate image-caption pairs keeping identical scores and references. The data augmentation strategy produces a low correlation in validation as there are plenty of pairs with close similarities and equal human scores.

	Composite	Flickr8k	VTT'19	VTT'20
	τ_c	τ_c	r	r
BLEU-4*	30.6	30.8	4.9	11.2
ROUGE-L*	32.4	32.3	15.8	20.9
METEOR*	38.9	41.8	22.2	23.5
CIDEr*	37.7	43.9	15.5	20.8
SPICE*	40.3	44.9	-	-
TIGEr*	45.4	49.3	-	-
BERTHA*	-	-	22.5	24.7
CLIP	**54.7**	46.2	23.3	**37.2**
CLOOB	49.9	**56.1**	**30.1**	34.9

Table 1. Flickr8k and Composite Kendall's correlation with human assesment. TRECVid-VTT 2019 and 2020 Pearson's correlation with human assesment. Only using the middle frame to simulated a image captioning dataset. *results from [11] for Flickr-8k and Composite and from [8] for TRECVid-VTT 2019 and 2020.

	VTT'19	VTT'20
	τ_c	τ_c
METEOR	16.1	23.5
CIDEr	13.9	20.8
CLIP_image	16.7	24.9
CLIP_similarity	16.2	26.9
CLIP_knn	17.0	**27.2**
CLIP_kmeans	16.7	27.1
CLOOB_image	19.6	22.4
CLOOB_similarity	20.4	23.5
CLOOB_knn	20.5	22.8
CLOOB_kmeans	**20.6**	23.4

Table 2. TRECVid-VTT Kendall's τ correlation with human assessment using frame data. Images is with a single image; similarity is computing an average between the frames

2.5. Filtering poor human judgments

In a dataset, there is always the possibility that one annotation does not follow the same criteria as another, even if it has come from a good annotator. We design a simple way of filtering some of the poor human judgments.

Pearson correlation is based on computing a linear correlation between the two sets. Spearman computes a monotonic function, more exactly it computes the Pearson correlation in the ranks. Both look for a linear function that fits either the data or the ranks. Considering these premises, the line that fits the sub-data being analysed is straightforward to be found. Then there are various techniques to find outliers to this line.

We use Cook's distance [17] as the measure to identify outliers. Cook's distances measures the influence of a data point on a fitted regression. A rule of thumb is to select four times the mean of all the distances as the distance in which a sample is classified as a outlier. We use this assumption to discard poor human judgments considering them as outliers.

3. EXPERIMENTAL RESULTS

3.1. Image captioning evaluation

The first step is to test both models in image captioning datasets and in a video captioning one, whereby as an initial experiment only the middle frame is used. Table 1 present the results together with some of the most popular metrics.

The datasets used for image captioning are Flickr 8k [18] consisting of 17k human evaluations of video and captions pairs with scores between 1 to 4, and Composite [19], which groups images from three different datasets: MSCOCO [20], Flickr8k [18], and Flickr30k [21]. The final dataset contains approx. 4k image-caption pairs with four human evaluations per pair. Kendall's correlation is used in both datasets. For video captioning, the last two years of TRECVid-VTT are used. Each year consists of less than 1541 videos-captiosn pairs from around ten systems. Each pair additionally have five references annotated by humans. Following the original paper [8] we use Pearson's correlation to compare human judgment to the metrics scores. However, we switch to Kendall's correlation in further experiments as it is the measure most used in the literature. Table 1 presents the video captioning results.

Across all the datasets, both methods obtain state of the art results. Neither sufficiently outperforms the other to declare it the best performing one. Each one gets good scores in different sets. Comparing the other metrics, we can see that these methods obtain a better correlation than the classical metrics. Only TIGEr has comparable performance. We can conclude that either the visual information is needed to get the final score or that we need to compare the caption to the references in a space where the sentences are related in a visual space. The embedding spaces are not built solely from a textual corpus but also a visual one, so sentences are closer if they appear in similar images instead of just in the context of a paragraph.

3.2. Video captioning evaluation

In this section, we discuss the application of this methodology in the video captioning scenario. The first thing to note is that none of these metrics has a obviously correct way to be applied in video evaluation. From analyzing individual videos, we observe modest differences in the similarity between the caption to other frames in the video. Table 2 compares the different ways to include this information.

The differences between mean/min/max are not included because they obtain similar correlations, all of them are represented as the similarity result.From the clustering experi-

	VTT'20
	τ_c
Trained CLIP	27.6 (+1.4%)
Trained CLOOB	22.9 (-2.5%)

Table 3. TRECVID training using TRECVid-VTT 2016 to 2019 as train/validation and test in 2020.

ments, we can see that k-NN and k-means perform similarly in these short videos, and they both obtain some of the best results. This result means that removing noisy frames from the video can improve the evaluation methodology and that the whole video can be summarized in a few keyframes. k-NN also shows that focusing on evaluating a few frames and measuring how the similarity changes leads to a better evaluation. This implies that only using a few frames to perform the evaluation is necessary and that the difficulty is in finding these frames.

3.3. Training results

One way to increase the performances in this task is to retrain the models in data from the task. In our setup we use TRECVid-VTT 2016 to 2019 [8] as training and validation data and test in TRECVid-VTT 2020. The text data is augmented using substitutions, remove, insertion, swap, and synonym replacement adding 10 new pairs from the original ones. The model is trained between 100 to 1000 epochs, with Adam using a learning rate of 10^{-8} and a weight decay of 0.001. The model with the highest Kendall's correlation in validation is used to test. The experiments took from 1 to 3 days to run om a NVIDIA RTX 2080TI.

Table 3 shows the final results for the best CLIP and CLOOB model. We see an improvement in CLIP while CLOOB does not perform as well as the original. This may be because the loss function, InfoLOOB, is more complex and may need additional tricks and data to perform well.

3.4. Filtering human annotations

In Table 4, we observe the effect of removing outliers from the annotations. We analyze both the percentage of captions and the improvement across the metrics. Table 4 shows the results from removing outliers detecting them using the actual model predictions and human judgments, Filter linear, or the ranks of both metrics, Filter ranks.

All the results are positive except one case with CIDEr. Assuming that the scores should follow a linear pattern and removing the outliers based on this assumption, we increase between 10% to 20% when only removing 6% to 7% of the captions. Assuming that the score should be linearly correlated with annotator scores in rank, we obtain substantially better results, achieving an increase of up to 30% in the metrics and only removing a 4% of the annotations. These experiments demonstrate that there are captions that penalize the correlations, meaning that these scores are not annotated following the same rules as the others and are outliers.

	VTT'20	# captions
	τ_c	remove
Filter linear CLIP	27.5 (+1.1%)	107 (6.9%)
Filter linear CLOOB	25.6 (+8.9)	93 (6%)
Filter linear CIDEr	20.7 (-0.4%)	86 (5.5%)
Filter linear METEOR	23.2 (+20.8)	106 (6.9%)
Filter rank CLIP	32.2 (+18.3%)	57 (3.6%)
Filter rank CLOOB	28.8 (+22.5%)	59 (3.8%)
Filter rank CIDEr	27.3 (+26.9%)	61 (3.9%)
Filter rank METEOR	25.3 (+31.7%)	56 (3.6%)

Table 4. TRECVid filtering captions base on linearity in the scores and on linarity in the ranks. Results use Kendall's correlation with human assesment and the percentatge increase with respect to the baseline. Also shown is the number of image-caption pairs removed and the percentage that this represents of the total number of samples.

4. CONCLUSION

We propose a new way to evaluate video captions using vision and language models. This can be accomplished without human annotations; however, the best performance is obtained using these references. One problem with these models is that they do not include the addition of multiple frames. To solve this problem we propose to use k-means or k-NN to find the best frames for the evaluation process. We retrain the model to adapt it better to the sentences used in video captioning. Finally, we also noticed some outliers in the datasets. Filtering these was shown to improve all the metrics without removing too many pairs.

Exploring better ways to include the video content can result in an improvement on the metric. Our initial ideas include better detection of relevant frames and/or adding the motion information to the evaluation if the captioning contains some action verb. The methods proposed to detect outliers can also be improved by explicitly detecting the human's bias. These considerations will be explore in future work.

5. ACKNOWLEDGMENTS

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

6. REFERENCES

[1] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.

[2] Chin-Yew Lin, "Rouge: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, 2004, pp. 74–81.

[3] Satanjeev Banerjee and Alon Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005, pp. 65–72.

[4] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh, "Cider: Consensus-based image description evaluation," in *CVPR*, 2015, pp. 4566–4575.

[5] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould, "Spice: Semantic propositional image caption evaluation," in *ECCV*. Springer, 2016, pp. 382–398.

[6] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi, "Bertscore: Evaluating text generation with bert," in *ICLR*, 2020.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *ACL*, 2019, pp. 4171–4186.

[8] Luis Lebron, Yvette Graham, Kevin McGuinness, Konstantinos Kouramas, and Noel E. O'Connor, "Bertha: Video captioning evaluation via transfer-learned human assessment," in *Proceedings of the Language Resources and Evaluation Conference*, Marseille, France, June 2022, pp. 1566–1575, European Language Resources Association.

[9] Ming Jiang, Qiuyuan Huang, Lei Zhang, Xin Wang, Pengchuan Zhang, Zhe Gan, Jana Diesner, and Jianfeng Gao, "Tiger: Text-to-image grounding for image caption evaluation," *arXiv preprint arXiv:1909.02050*, 2019.

[10] Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge Belongie, "Learning to evaluate image captioning," in *CVPR*, 2018, pp. 5804–5812.

[11] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi, "CLIPScore: a reference-free evaluation metric for image captioning," in *EMNLP*, 2021.

[12] Andreas Fürst, Elisabeth Rumetshofer, Viet Tran, Hubert Ramsauer, Fei Tang, Johannes Lehner, David Kreil, Michael Kopp, Günter Klambauer, Angela Bitto-Nemling, et al., "Cloob: Modern hopfield networks with infoloob outperform clip," *arXiv preprint arXiv:2110.11316*, 2021.

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[14] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu, "Uniter: Universal image-text representation learning," in *ECCV*, 2020.

[15] Evelyn Fix and Joseph Lawson Hodges, "Discriminatory analysis. nonparametric discrimination: Consistency properties," *Int Stat Rev*, vol. 57, no. 3, pp. 238–247, 1989.

[16] Stuart Lloyd, "Least squares quantization in pcm," *IEEE Trans Inf Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[17] R Dennis Cook, "Detection of influential observation in linear regression," *Technometrics*, vol. 19, no. 1, pp. 15–18, 1977.

[18] Micah Hodosh, Peter Young, and Julia Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *IJCAI*, vol. 47, pp. 853–899, 2013.

[19] Somak Aditya, Yezhou Yang, Chitta Baral, Cornelia Fermuller, and Yiannis Aloimonos, "From images to sentences through scene description graphs using commonsense reasoning and knowledge," *arXiv preprint arXiv:1511.03292*, 2015.

[20] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.

[21] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2641–2649.