2022 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-9620-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICIP46576.2022.9898011

UCF-CAP, VIDEO CAPTIONING IN THE WILD

Christos Chatzikonstantinou, Georgios Grigorios Valasidis, Konstantinos Stavridis,
Georgios Malogiannis, Apostolos Axenopoulos and Petros Daras, Senior Member IEEE

The Visual Computing Lab, Information Technologies Institute, Centre for Research and Technology Hellas
{chatziko,valasidis,staurid,giorgosmalo,axenop,daras}@iti.gr

ABSTRACT

Recent technological advances in the fields of data and computer science have improved significantly the everyday life of people. However, technological advances are also being adopted by criminals to facilitate and expand their illicit actions. The Deep Learning (DL) paradigm has shown a significant potential in analysing complex structured data. However, in the crime detection domain, a limited number of public datasets is available, constrained to specific tasks only, which hinders the research and development of accurate and robust DL-assisted tools. The goal of this work is to extend the well-known UCF-crime dataset to the case of video captioning. To the best of our knowledge, this is the first publicly available crime-related video captioning dataset. A new proposed video captioning approach is compared to a plethora of state-of- the-art-methods in this dataset, while qualitative and quantitative characteristics of the latter are presented.

Index Terms— captioning dataset, crime, transformer

1. INTRODUCTION

Living in the era of the evolution of the WWW, a plethora of human activities (personal, commercial, professional, educational etc.) have been reformed to be available online. Regarding the pandemic situation, the rate of crimes that are using WWW as part of the whole process or being committed exclusively online has been dramatically increased due to the fact that the candidate victims used to spend a lot of time connected. Therefore, it is of great importance to develop robust tools, based on Artificial Intelligence (AI), capable of detecting, preventing and combating crime on the internet [1]. In order to benefit from the DL paradigm for creating such tools, large volumes of annotated data are required, while the open-research approach must be followed to attract as many researchers as possible.

Most of the crime-related publicly available datasets, refer to past crime record data published from several Law Enforcement Agencies (LEA) mainly located in the United States of America. However, it is very significant for LEAs to be equipped with robust tools capable of detecting, from content that is published from perpetrators on the internet, currently organized or under execution crimes.

Following the above line of investigation, the UCF-crime dataset [2] was introduced to provide annotations for anomaly detection and event classification (shooting, arson, stealing, etc) in videos. In complex automated video understanding tasks, such as video captioning, creating annotations to train efficient DL algorithms is highly challenging. The video captioning task enables the automated production of the description of a video in natural language text in order to be easily read by a LEA officer. Moreover, text allows further semantic processing.

Several publicly available datasets have been developed in the context of video captioning. They can be classified into three main categories: a) Classic video captioning datasets that each video is described by one sentence (MSVD [3], MSRVTT [4], MPII-MD [5], M-VAD [6], Charades [7] and VTW [8]); b) Dense video captioning datasets that describe a video using a set of sentences (paragraph), each sentence is temporally localized with a timestamp (TACoS [9], YouCook2 [10], VideoStory [11] and ActivityNet Caption [12]), and c) Grounded video captioning datasets (ActivityNet Entities [13]), that describe videos by one sentence that is grounded on the visual objects. Regarding the domain of the video captioning datasets, most of them refer to cooking [14, 15], movies [5, 6] and social media videos.

The main goal of the current work is to fill the gap of appropriate video captioning data in the crime domain. The main contributions of this study are summarized as follows:

- Introduction of the UCF-CAP video captioning dataset. Natural language descriptions for a subset of UCF-Crime videos are provided by adopting the classic video captioning paradigm, comparable in size with well-established datasets in the field. The UCF-CAP is publicly available (`https://zenodo.org/record/6821915#.Ys1st-xBwi0`).

- Propose a novel DL-based video captioning algorithm that adapts optimally to the introduced dataset by capturing fruitful correlation among words and visual information.

The work presented in this paper was supported by the European Commission under contract H2020-883596 AIDA.

- Experimental evaluation of the proposed algorithm against a variety of competitive state-of-the-art methods on the new dataset.

The rest of the study is organised as follows: Section 2 introduces the UCF-CAP dataset. The proposed DL-based video captioning algorithm is presented in Section 3, the evaluation scheme is presented in Section 4 and the paper is concluded in Section 5.

2. THE INTRODUCED DATASET

In this section, the UCF-CAP dataset is presented along with the UCF-Crime dataset that was employed as the visual stimulus of the captioning task.

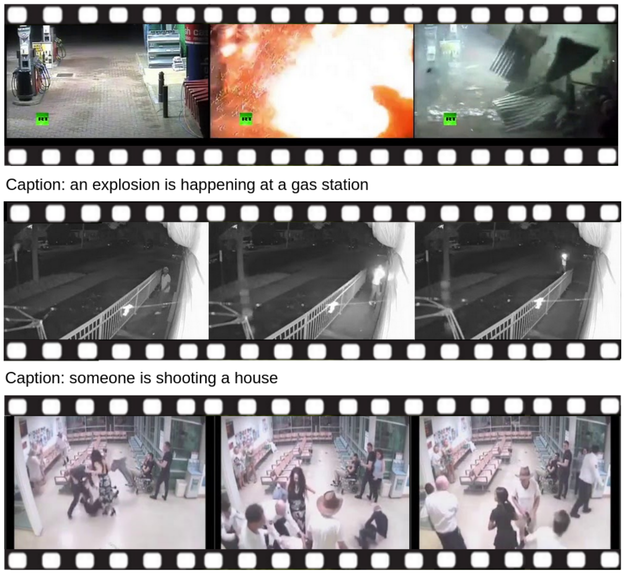
2.1. UCF-Crime

UCF-Crime [2] was originally created to deal with the task of video-based anomaly detection. It contains long, untrimmed videos from CCTV cameras covering 13 real-world anomalies, including Abuse, Arrest, Arson, Assault, Burglary, Explosion, Fighting, Road Accident, Robbery, Shooting, Shoplifting, Stealing and Vandalism, as well as videos with Normal activities. Specifically, there are 950 videos with at least one anomaly event, plus 950 videos without any anomaly.

2.2. UCF-CAP

For the task of video captioning within the crime-related domain, a single caption per video is assigned. All anomaly videos are divided into clips with duration up to 20 seconds, based on the intuition that crime activities are usually short, as illustrated in Figure 1. The clips are extracted from the UCF-Crime dataset videos and 2012 of them are randomly assigned to four male annotators, with average age 35 years, to produce captions. Clips that show nothing but a logo are excluded from annotation. All clips are described by English sentences in simple present, present continuous or present perfect [simple] tenses. The dataset is split according to 70%:10%:20% scheme corresponding to 1418, 198 and 396 clips in training, validation and testing sets, respectively. Table 2 shows the number of used clips per action class, with respect to the subset that they belong to.

The complete dataset contains captions for clips of 862 different videos out of the 950 with anomalies, 2012 sentences and a vocabulary of 888 words (analyzed as: Nouns-606, Verbs-177, Adjectives-32, Adverbs-17, etc.). Each clip is annotated with one sentence including, on average, 8.28 words. Table 1 presents the statistics of UCF-CAP in comparison with MSVD [3] and MSR-VTT [4] datasets. Figure 2 shows the proportion of most used verbs (activity indicators), where the verbs that denote a crime-related activity are pointed out.



Caption: an explosion is happening at a gas station

Caption: someone is shooting a house

Caption: people are fighting at a room while three security guards interfere

Fig. 1: Three caption examples.

3. PROPOSED VIDEO CAPTIONING METHOD

The proposed video captioning method is based on the baseline of the transformer-based method presented in [16] to better capture correlations in the spatio-temporal space from a constrained in size dataset. Two modalities are employed as transformer inputs to enhance the efficiency: RGB features extracted by a 2D CNN and motion features extracted by a 3D CNN.

The complexity of the proposed method is analogous to the transformer model. The model consists of: convolutional layers ($(O(k * n * d^2))$, multi-head attention layers ($O(n^2 * d + n * d^2)$ and fully connected layers ($O(n^2)$), n denotes the sequence length, d is the number of filters in convolutional layers and the dmodel dimension in multihead attention layers. The kernel size of convolutions is denoted as k and it can be omitted in convolutional layer computations as it is much smaller compared to n and d. Thus the total complexity is: $C = O(n*d^2) + O(n^2*d + n*d^2) + O(n^2) = O(n*d^2) + O(n^2*d) + (n*d^2) + O(n^2) = O(2*n*d^2 + n^2*d)$ $(d >> 1)$.

Concerning the fusion of the above modalities, different fusion schemes were tested (illustrated also in Figure 3a):

- Fusion scheme A: Both modalities are fused before the Transformer's encoder

- Fusion scheme B: Each modality is processed by a different encoder and they are fused before the decoder

Regarding the fusion operator, different forms are also evaluated, as illustrated in Figure 3b. More specifically:

i) An addition of the modalities

Dataset	Domain	Videos	avg length	# clips	# sent	# sents/clip	# words	vocab	# verbs	unique verbs	len (hrs)
MSVD [3]	open	1,970	10 sec	1,970	70,028	41	607,339	13,010	114,878	2,268	5.3
MSR-VTT [4]	open	7,180	20 sec	10,000	200,000	20	1,856,523	29,316	270,564	4,105	41.2
UCF-CAP	crime	862	9.96 sec	2,012	2,012	1	16,658	888	2,480	177	5.5

Table 1: Dataset's comparison.

action-class	train	validation	test
Abuse	101	14	28
Arrest	101	14	28
Arson	101	14	28
Assault	101	14	28
Burglary	129	18	36
Explosion	100	14	28
Fighting	101	14	28
Road Accidents	129	18	36
Robbery	127	18	36
Shooting	101	14	28
Shoplifting	101	14	28
Stealing	127	18	36
Vandalism	99	14	28

Table 2: Number of used clips per subset for each action class.

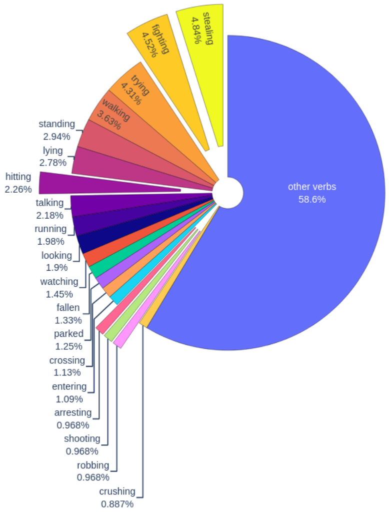


Fig. 2: Pie chart of used verbs.

ii) The addition of the output of two multi-head attention layers

iii) A concatenation of the modalities

iv) The concatenation of the output of two multi-head attention layers

Using the multi-head attention layers the model can simultaneously handle the information from different representation sub-spaces at different positions. Thus, the input of each multi-head attention layer consists of information from both modalities for an enriched representation.

4. EXPERIMENTAL EVALUATION

The proposed video captioning method is used for the evaluation of the introduced UCF-CAP dataset and thus forming an initial basis for the video captioning task in the crime domain. In particular, an ablation study is presented in order to test the performance of the proposed method for different modalities. Moreover, our video captioning method is compared, on the UCF-CAP dataset, to state-of-the-art video captioning methods. The code is available at `https://gitlab.com/chatzikon/ucf_cap`.

4.1. Evaluation metrics

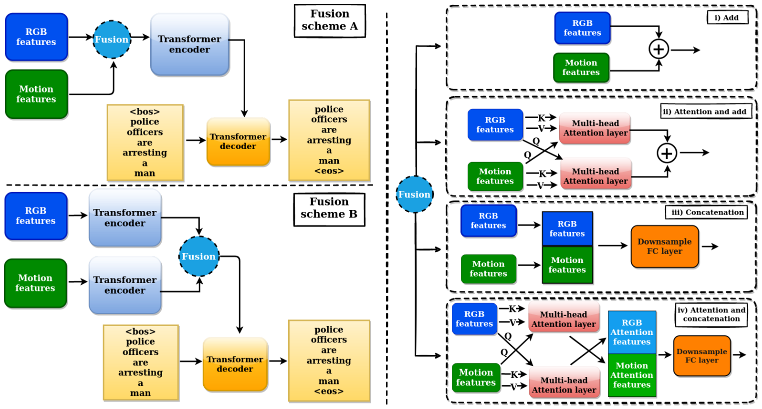
For the evaluation of this study, three common metrics are employed to evaluate the quality of the text generated by a DL model. Those are: METEOR, ROUGE-L and CIDEr. The Microsoft COCO Evaluation Server[17] is used to compute the above metrics.

4.2. Implementation details

To train, validate and test the proposed method, the following experimental schemes are employed. The train, validation and test splits presented in Section 2.2 are used. During the experiments, the Python 3.8 and Pytorch (version 1.5.1) environments are adopted.

Regarding the transformer hyperparameters, a batch size of 10 is adopted, the number of encoder/decoder layers are set to 9, the threshold for gradient clipping is set to 0.75 and the maximum sentence length is set to 20. The initial learning rate is equal to 0.01 and it is divided by two if the loss is not reduced after 6 epochs. The sgd optimizer is employed. The dimension of the model, d_{model} is 1024 and the hidden state size of the feed-forward layer is set as 2048. The model is trained for 100 epochs and greedy decoding is employed for inference.

Concerning the feature extraction, a ResNet152 model is used for the extraction of the RGB features and a C3D model for the extraction of the motion features. In both cases, the interval between successive frames or clips is 5 frames. Furthermore, concerning the motion features, each clip consists of 16 consecutive frames. Since the videos of the dataset do not have equal size, the maximum number of frames to ex-



(a) Fusion schemes

(b) Fusion operator forms

Fig. 3: Fusion schemes employed in this study

tract features is set to 50. If a video is smaller, a zero padding process is performed.

4.3. Results

The ablation study is performed on the multimodal transformer presented on Section 3. The different schemes presented on Section 3 are evaluated and the results are illustrated in Table 3. In this Table the results of a unimodal model, employing only RGB features, are also presented.

The best results are achieved from the model using the fusion scheme A with attention and addition and the model using the fusion scheme B with addition. Thus, it can be concluded that the addition of different modalities leads to better results than their concatenation. Moreover, those models perform better compared to the unimodal model, validating the superiority of the multimodal approach compared to the unimodal one.

Fusion scheme	METEOR	ROUGE-L	CIDEr
Unimodal	12.19	**29.69**	58.15
A-i	12.61	29.40	55.96
A-ii	**13.23**	**30.09**	**63.63**
A-iii	12.85	28.99	56.27
A-iv	12.06	28.29	45.18
B-i	**13.03**	29.08	**68.53**
B-ii	12.08	27.22	49.52
B-iii	12.7	29.63	54.76
B-iv	12.46	29.26	56.03

Table 3: Impact of different fusion schemes on the performance of the video captioning model.

Finally, the best results of the multimodal transformer are compared to the results of four recent state-of-the-art models, the ARB model of [18], the SGN model of [19], the SAVC model of [20] and the RecNet model of [21]. All models are multimodal architectures, employing RGB and motion features, except from the RecNet model that is a unimodal architecture. The results can be seen in Table 4. The proposed multimodal transformer architecture achieves better results showing that a multimodal multi-layer transformer architecture can better capture the dependencies of the visual and textual information, compared to the coarse-to-fine procedure of [18], the semantic guided approach of [19] and [20] or the reconstruction approach of [21] . Based on the above experimental scheme, it can be derived that the proposed video captioning method is well evaluated on the introduced dataset and it can serve as a robust baseline for future studies.

Fusion scheme	METEOR	ROUGE-L	CIDEr
A-ii	**13.23**	**30.09**	**63.63**
B-i	**13.03**	29.08	**68.53**
ARB [18]	10.16	27.66	26.61
SGN [19]	12.03	29.93	58.40
SAVC [20]	10.37	27.11	23.08
RecNet [21]	13.01	**30.29**	52.18

Table 4: Comparative evaluation of SoTA video captioning methods employing the introduced dataset.

5. CONCLUSIONS

In this paper the first crime-related video captioning dataset was introduced. Moreover, a customized DL algorithm is proposed that optimally correlates the textual and video information of the dataset, based on a multimodal transformer architecture, fusing RGB and motion information. Directions of future work include the enrichment of the information provided by the dataset such as the description of each video with a set of sentences, including temporally localized timestamps, or the grounding of each annotation sentence on the video objects.

6. REFERENCES

[1] Panagiotis Stalidis, Theodoros Semertzidis, and Petros Daras, "Examining deep learning architectures for crime classification and prediction," *Forecasting*, vol. 3, no. 4, pp. 741–762, 2021.

[2] Waqas Sultani, Chen Chen, and Mubarak Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6479–6488.

[3] David Chen and William B Dolan, "Collecting highly parallel data for paraphrase evaluation," in *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, 2011, pp. 190–200.

[4] Jun Xu, Tao Mei, Ting Yao, and Yong Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5288–5296.

[5] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele, "A dataset for movie description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3202–3212.

[6] Atousa Torabi, Christopher Pal, Hugo Larochelle, and Aaron Courville, "Using descriptive video services to create a large data source for video annotation research," *arXiv preprint arXiv:1503.01070*, 2015.

[7] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *European Conference on Computer Vision*. Springer, 2016, pp. 510–526.

[8] Kuo-Hao Zeng, Tseng-Hung Chen, Juan Carlos Niebles, and Min Sun, "Generation for user generated videos," in *European conference on computer vision*. Springer, 2016, pp. 609–625.

[9] Marcus Rohrbach, Michaela Regneri, Mykhaylo Andriluka, Sikandar Amin, Manfred Pinkal, and Bernt Schiele, "Script data for attribute-based recognition of composite activities," in *European conference on computer vision*. Springer, 2012, pp. 144–157.

[10] Luowei Zhou, Chenliang Xu, and Jason J Corso, "Towards automatic learning of procedures from web instructional videos," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[11] Spandana Gella, Mike Lewis, and Marcus Rohrbach, "A dataset for telling the stories of social media videos," in *Proceedings of the 2018 conference on empirical methods in natural language processing*, 2018, pp. 968–974.

[12] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles, "Dense-captioning events in videos," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 706–715.

[13] Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J Corso, and Marcus Rohrbach, "Grounded video description," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6578–6587.

[14] Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele, "A database for fine grained activity detection of cooking activities," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 1194–1201.

[15] Pradipto Das, Chenliang Xu, Richard F Doell, and Jason J Corso, "A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2634–2641.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[17] 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.

[18] Bang Yang, Yuexian Zou, Fenglin Liu, and Can Zhang, "Non-autoregressive coarse-to-fine video captioning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 3119–3127.

[19] Hobin Ryu, Sunghun Kang, Haeyong Kang, and Chang D Yoo, "Semantic grouping network for video captioning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 2514–2522.

[20] Haoran Chen, Ke Lin, Alexander Maye, Jianmin Li, and Xiaolin Hu, "A semantics-assisted video captioning model trained with scheduled sampling," *Frontiers in Robotics and AI*, vol. 7, pp. 475767, 2020.

[21] Bairui Wang, Lin Ma, Wei Zhang, and Wei Liu, "Reconstruction network for video captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7622–7631.