Video description by textual and semantic enrichment

Keywords: video description, deep learning, convolutional neural network, coreference resolution, knowledge graph, multi-tasking, multimodality

Context and motivation

Video description methods aim at automatically generating textual descriptions from video sequences/clips. This is a well-known problem in the field of computer vision where many methods have been proposed in the literature [1]. This topic is related to performance evaluation and in particular to databases. Many databases have been proposed for performance evaluation of video description methods [2]. The two most widely used are the MSVD [3] and MSR-VTT [4] databases. These databases are usually built from web videos (e.g. Youtube) on targeted categories (e.g. cooking, e-commerce, social networks) or in multi-category. Textual annotations are most often produced by hand via web services like Amazon Mechanical Turk.

Nevertheless, the databases proposed in the literature have severe limitations [2]. The encoding/quality of the videos is heterogeneous and the videos are provided without multimodality (no or few audio tracks, subtitles and metadata). They are not scaled and the test sets are provided in black box mode. Finally, the descriptions are given a posteriori (without temporality) and are limited to visual information without contextualization (e.g. which person, which place, which date, etc.). As discussed in [1, 2], the constitution of scaled, standardized databases with textual information (audio tracks, subtitles and metadata), structured in test sets and offering a large ground truth is essential for future research on video description methods.

These limitations can be partly overcome by using television (TV) capture. Indeed, TV video streams, unlike Web video streams, allow a standardized capture, at scale and with textual information. Various works have been carried out in the past for the constitution of TV databases for the detection of video segments [5], data journalism [6] and natural language processing [7]. Nevertheless, automatically and reliably generating video descriptions that meet ground truth requirements remains an open issue. Even if video description methods have gained a lot of maturity (and in particular those based on deep learning [1]), they can neither guarantee the level of robustness necessary for the elaboration of a ground truth, nor cross (on the sole basis of visual analysis) the semantic gap necessary for contextualization.

Objective ot the thesis

In order to address this problem, we propose in this PhD thesis to explore natural language processing approaches for semantic representation and deep learning in order to propose a new system for contextual video description. The idea is to map the textual data contained in the electronic program guides to the knowledge graphs of the Web of Data, and then to link these data to the video sequences associated with them thanks to the subtitles and their timestamp. At the textual level, the link between the different modalities raises several scientific challenges such as the resolution of co-references [8, 9] between different natures of texts (written, oral transcription) and multilingual processing [10]. On video analysis, multi-task [11, 12] and multi-modal processing [13, 14] will have to be proposed for performance evaluation [5, 6] for the generation of the database and the ground truth.

This work is therefore at the crossroads between deep learning on videos and NLP based on knowledge graphs. The main expected contributions are :

- Proposal of a method to generate a knowledge graph from the processing of audio descriptions
- Proposal of a deep learning model on video descriptions integrating a knowledge graph for the generation of summaries
- Production of a database of video descriptions

Context and supervision

This PhD is a full funded position supported with a three years grant. The student will be hosted by the LIFAT laboratory[1] of the University of Tours[2], this research direction extending and reinforcing collaborations already in progress between the BDTLN and the RFAI[3] teams. The work will be supervised by Arnaud Soulet (Assistant Professor), Donatello Conte (Full Professor) and co-supervised by Nathalie Friburger (Assistant Professor), Mathieu Delalandre[4] (Assistant Professor). Applications have to be submitted online on the ADUM platform[5] before May 15,2023.

Link to apply

References

[1] S. Li and al. Visual to Text: Survey of Image and Video Captioning. Transactions on Emerging Topics in Computational Intelligence, vol. 3(4), pp. 297-311, 2019.
[2] M. Rafiq and al. Video Description: Datasets & Evaluation Metrics. IEEE Access, vol. 9, 2021.
[3] D.L. Chen and W.B. Dolan. Collecting highly parallel data for paraphrase evaluation. Annual Meeting of the Association for Computational Linguistics: Human Language Technologie, pp. 190-200, vol. 1, 2011.
[4] J. Xu and al. MSR-VTT: A large video description dataset for bridging video and language. Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5288–5296, 2016.
[5] V.H. Le and al. A large-Scale TV Dataset for partial video copy detection. International Conference on Image Analysis and Processing (ICIAP), Lecture Notes in Computer Science (LNCS), vol 13233, pp. 388-399, 2022.
[6] F. Rayar and al. A large-scale TV video and metadata database for French political content analysis and fact-checking. Conference on Content-Based Multimedia Indexing (CBMI), 2022.
[7] P. Lison and J. Tiedemann. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. Language Resources and Evaluation Conference (LREC), 2016.
[8] R. Sukthanker, S. Poria, E. Cambria, R. Thirunavukarasu, Anaphora and coreference resolution: A review, Information Fusion, Volume 59, 2020, Pages 139-162, ISSN 1566-2535
[9] V. Ramanathan, A. Joulin, P. Liang and L. Fei-Fei, Linking People in Videos with ``Their'' Names Using Coreference Resolution, in Computer Vision -- ECCV 2014, 2014, Springer International Publishing, pp. 95-110
[10] Oliveira, I.L., Fileto, R., Speck, R., Garcia, L.P., Moussallem, D. and Lehmann, J., 2021. Towards holistic entity linking: Survey and directions. Information Systems, 95, p.101624.
[11] Z. Liu and al. Multi-Task Video Captioning with a Stepwise Multimodal Encoder. Electronics, vol. 11(17), pp. 2639, 2022.
[12] S. Chen and al. Video Captioning with Guidance of Multimodal Latent Topics. International Conference on Multimedia (MM), pp. 1838–1846, 2017.
[13] D. Ramachandram and G. W. Taylor. Deep Multimodal Learning: A Survey on Recent Advances and Trends. Signal Processing Magazine, vol. 34 (6), pp. 96-108, 2017.
[14] J. Summaira and al. Recent Advances and Trends in Multimodal Deep Learning: A Review. arXiv.2105.11087, 2021.

[1] https://lifat.univ-tours.fr/
[2] https://international.univ-tours.fr/
[3] https://www.rfai.lifat.univ-tours.fr/
[4] http://mathieu.delalandre.free.fr/
[5] https://adum.fr/