Video captioning: describing our behaviors with a few words

Type: Research internship in Computer Vision

Level: Master 1, Master 2, Fourth or Fifth-year Engineer

Supervision:

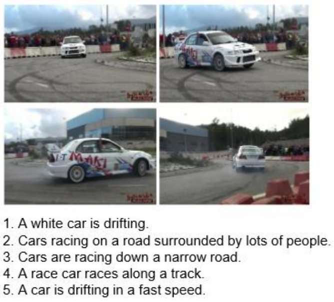
- Carlos Crispim-Junior (Associate Professor)
- Laure Tougne (Full professor)

Location: LIRIS UMR CNRS 5205, Université Lumière Lyon 2, Bâtiment C, 5 avenue Pierre Mendès-France, 69676 Bron cedex

Keywords: deep learning, action detection, self-driving cars, in-cabin activity analysis

Context: Self-driving cars (SDC) have gained significant attention since the progress of artificial intelligence for visual scene understanding in the 2010s. Technology companies such as Waymo and Tesla now compete with established car manufacturers in the development of SDCs. Cruise and Waymo have already deployed robot taxi services in a few cities in the United States. However, research on what will be the typical activities of the occupants of an SDC is still an open problem, particularly in vehicles of automation levels 3 and 4 (SAE[1] standard). In this context, the AURA AutoBehave project (2019-2023) seeks to develop methods to automatically analyze the activities of passengers of SDCs, and to study how the changes brought by the usage of SDCs may influence our lives in terms of in-vehicle postures and actions. Changes in such behavioral patterns may affect our in-vehicle comfort, security, and how we value our travel time.

Subject: In video captioning (or description, Figure 1) research, we seek to develop methods that can describe the content of a sequence of images in textual form. Early work has focused on describing a video clip with a single sentence, but as the field progressed, research has targeted the generation of denser and semantically finer descriptions (Krishna et al., 2017; Xiujun et al., 2020; Estevam et al., 2021). Applications of such methods range from automatic generation of video subtitles to assistive technology, like scene description for visually impaired people, or road description for driver assistance systems. During this internship, we will study the recent methods, datasets, and evaluation measures used by the state-of-the-art approaches, try to replicate their performance by following a common, reproducibility approach, and then defy the performance of the selected method by evaluating them in real-world, limited-size datasets acquired in the context of AutoBehave dataset.



1. A white car is drifting.
2. Cars racing on a road surrounded by lots of people.
3. Cars are racing down a narrow road.
4. A race car races along a track.
5. A car is drifting in a fast speed.

Figure 1. Example of Video description output, source: Xu et al, 2016

[1] Society of Automotive Engineers

Tasks

- Revise the state of the art on methods for video description
- Identify current datasets in use and the metrics to evaluate their performance
- Draw a short list of relevant methods to construct a new benchmark
- Evaluate the performance of selected methods in the AutoBehave dataset, both quantitatively and qualitatively
- Write a report in article form to describe the work carried out

Expected skills

- Python programming
- PyTorch programming and OpenCV library skills will be considered a plus

Profile of the candidate:

We are looking for a motivated candidate with a strong background in computer science and applied mathematics. Experience in image processing, computer vision, and/or machine learning will be a plus. The intern will have the opportunity to collaborate on the writing of a research article about the work realized to be submitted to a major conference in the field of computer vision.

Required skills:

- Language Python
- OpenCV library

The following skills would be counted as a plus:

- Versioning tools (GIT)
- Framework PyTorch or TensorFlow.

Starting date: February/March 2024

Salary: “gratification de stage” in France

Contact : carlos.crispim-junior@liris.cnrs.fr

References

1. Jain, Vanita, Fadi Al-Turjman, Gopal Chaudhary, Devang Nayar, Varun Gupta, et Aayush Kumar. « Video Captioning: A Review of Theory, Techniques and Practices ». Multimedia Tools and Applications 81, n° 25 (1 October 2022): 35619-53. https://doi.org/10.1007/s11042-021-11878-w.
2. Xu, Jun, Tao Mei, Ting Yao, et Yong Rui. « MSR-VTT: A Large Video Description Dataset for Bridging Video and Language », 1 juin 2016. https://www.microsoft.com/en-us/research/publication/msr-vtt-a-large-video-description-dataset-for-bridging-video-and-language/.
3. Zhu, Wanrong, Bo Pang, Ashish V. Thapliyal, William Yang Wang, et Radu Soricut. « End-to-end Dense Video Captioning as Sequence Generation ». In Proceedings of the 29th International Conference on Computational Linguistics, 5651-65. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, 2022. https://aclanthology.org/2022.coling-1.498.
4. Zhu, Wanrong, Bo Pang, Ashish V. Thapliyal, William Yang Wang, et Radu Soricut. « End-to-end Dense Video Captioning as Sequence Generation ». In *Proceedings of the 29th International Conference on Computational Linguistics*, 5651-65. Gyeongju, Republic of Korea:

International Committee on Computational Linguistics, 2022. https://aclanthology.org/2022.coling-1.498.

5. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., & Carlos Niebles, J. (2017). Dense-captioning events in videos. In Proceedings of the IEEE international conference on computer vision (pp. 706-715).
6. Estevam, V., Laroca, R., Pedrini, H., & Menotti, D. (2021). Dense video captioning using unsupervised semantic information. arXiv preprint arXiv:2112.08455.
7. Li, Xiujun, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, et al. « Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks ». arXiv, 25 July 2020. https://doi.org/10.48550/arXiv.2004.06165.