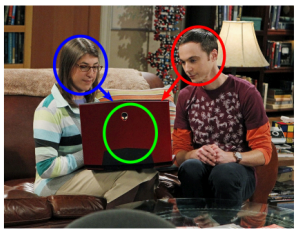
Internship: multimodal conversation script generation

Context

Face to face conversation remains the most natural means of communication between humans. It conveys much richer messages than typical text representations such as meeting minutes, thanks to the diversity and range of signals that can be passed through multimodal channels, such as emotions, visual grounding, etc. Handling such features remains elusive for machines, because of fundamental limitations in how multimodal understanding is implemented. Yet, latest developments in the field of AI show promise in handling audio-video captures of participants, and reasoning on the content of such inputs.

In order to address the gap in multimodal understanding of conversations, The MINERAL ANR project aims at generating enriched representations of multiparty conversations in the form of a conversation script similar to a movie script or a play script. These representations will include transcripts of the uttered speech with addressee, goal and communicative act, as well as textual descriptions of the activities and stances of each speaker, and their interactions with real-world objects. An imaginative goal is that actors should be able to replay the conversation from the script as they do with movies. Latent representations uncovered from performing this task are expected to enhance the understanding capabilities of AI models and allow for novel applications, such as generation of audio-visual summaries of face-to-face meetings.



[Amy and Sheldon are speaking to Amy's mother over a video computer connection on a laptop]
Mrs. Fowler : It's nice to meet you too, Sheldon. I honestly didn't believe Amy when she told me she had a boyfriend. *(to Sheldon Cooper)*
Sheldon Cooper : I assure you I am quite real. And I'm having regular intercourse with your daughter. *(to Mrs. Fowler)*
Mrs. Fowler : *[in a surprised tone]* What ?
Sheldon Cooper : Oh, yes. We're like wild animals in heat. It's a wonder neither of us has been hurt. *(to Mrs. Fowler)*
Mrs. Fowler : *[scared]* Amy, what is he saying ? *(to Amy Farrah Fowler)*
Amy Farrah Fowler : You wanted me to have a boyfriend, mother. Well, here he is. *(to Mrs. Fowler)*
[Sheldon waves at the computer screen, while Mrs. Fowler nervously waves back]

This internship is framed in the MINERAL ANR project, and aims at building an evaluation framework for assessing the quality of generated scripts based on existing movie and episode script datasets, and construct baselines for script generation with appropriate specialized models for underlying tasks such as scene description or transcript generation, plugged into large language models.

Objectives

The goal of this internship is twofold:

1) Propose an evaluation methodology for assessing the quality of a generated script
2) Build and assess baselines leveraging disjoint building blocks such as speech transcription and automated description of video scenes

Work resulting from the internship will be published in appropriate conferences and journals.

Work plan

The intern will first review methods for generating and evaluating textual representations from videos in the subfield of conversation analysis and summarization. The goal is to get a good understanding of current research problems and potential solutions.

Then, the next step consists in preparing existing datasets for the script generation task. Targeted datasets include minutes from debates at the assemblée nationale (including transcripts, reactions, stance, from audio and video), as well as the Bazinga TV series corpus including original scripts from episodes of Big Bang Theory, and manually structured transcripts.

Finally, the intern will implement baseline systems from existing feature extraction models such as OpenFace for face dynamics (expression, gaze…), Whisper for transcripts, Pyannote for speaker diarization, etc that will be fed to finetuned LLMs in textual form. If time permits, the intern will look into audio and video tokenizers, such as SpeechTokenizer and Cosmos-Tokenizer, in order to adapt the LLMs to raw features.

Practicalities

The internship will be funded ~500 euros per month for a duration of 6 months. It will take place in Marseille within the TALEP research group at LIS/CNRS on the Luminy campus. The intern will collaborate with other interns from the ANR project (at LISN and Orange Labs), as well as PhD students and researchers from the research group. A potential PhD funding on the same topic is also available in the project.

How to apply: send an application letter, transcripts and CV to benoit.favre@univ-amu.fr

- Application deadline: December 15th, 2024
- Expected start: early spring 2025

References

Soldan, M., Pardo, A., Alcázar, J. L., Caba, F., Zhao, C., Giancola, S., & Ghanem, B. (2022). Mad: A scalable dataset for language grounding in videos from movie audio descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5026-5035). https://github.com/Soldelli/MAD

Banchs, R. E. (2012, July). Movie-DiC: a movie dialogue corpus for research and development. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 203-207).

Lerner, P., Bergoënd, J., Guinaudeau, C., Bredin, H., Maurice, B., Lefevre, S., ... & Barras, C. (2022, June). Bazinga! a dataset for multi-party dialogues structuring. In *13th Conference on Language Resources and Evaluation (LREC 2022)* (pp. 3434-3441).

Baltrušaitis, T., Robinson, P., & Morency, L. P. (2016, March). Openface: an open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision (WACV)* (pp. 1-10). IEEE.

Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., ... & Gill, M. P. (2020, May). Pyannote. audio: neural building blocks for speaker diarization. In *ICASSP 2020-2020*

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7124-7128). IEEE.
Huang, B., Wang, X., Chen, H., Song, Z., & Zhu, W. (2024). Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14271-14280).
Wu, S., Fei, H., Qu, L., Ji, W., & Chua, T. S. (2023). Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*.
Ryoo, M., Piergiovanni, A. J., Arnab, A., Dehghani, M., & Angelova, A. (2021). Tokenlearner: Adaptive space-time tokenization for videos. *Advances in neural information processing systems*, *34*, 12786-12797.
Bain, M., Huh, J., Han, T., & Zisserman, A. (2023). Whisperx: Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv:2303.00747*.