Offer #2024-07408

PhD Position F/M Explainable and frugal audio scene description

Contract type : Fixed-term contract

Level of qualifications required : Graduate degree or equivalent

Fonction : PhD Position

Context

Inria Défense&Sécurité (Inria D&S) was created in 2020 to federate Inria's actions for the benefit of military forces. The PhD will be carried out within the audio processing research team of Inria D&S, under the supervision of Jean-François Bonastre and co-supervised by Raphaël Duroselle.

The automatic audio scene description task is to present operators with a summary of the information present in the scene, in the form of augmented text. This text provides a visual summary of the most important information, while efficiently structuring access to specific information. Here is an illustrative example of a summary: « This five-minute recording features three different speakers. Speaker A corresponds to a known identity in the database and speaks French with a strong Monawa accent, speakers B and C are unknown in the database and speak English in their interactions with A and use an unidentified language when talking to each other. The voices of B and C show strong similarities with speakers from the Eastern Quabar region. The main theme of the recording concerns a transfer of goods between the cities of Orienta and Flagrance. The date July 8, 2023 is mentioned three times.». Clicking on A gives the operator information about A and details of the voice identification performed. There will be direct access to the time segments during which A spoke and to their transcription. The transcription will highlight names of people, places or dates (named entities).

Assignment

Goal

The aim of this thesis is to propose a general framework for processing audio recordings for intelligence purposes. It consists in defining a high-level application adapted to the needs of end users, favouring the presentation of a recording in the form of a summary report to highlight its salient points.

Approach

This approach is inspired both by textual description of video scenes [1] and by dialogue systems based on audio-visual scenes [2]. The system will be based on the extraction of speech signal representations at different scales (frame, speech segment or sound event, complete recording), possibly dedicated to different tasks. The representations, useful for the various technological bricks of the system, will be embeddings extracted from deep neural networks, either generic [3] or dedicated to each task. The fusion between the different levels of information can be achieved with an architecture inspired by the multi-stream "Encoder-Decoder" scheme [4], with several encoders producing sequences of representations and one or more decoders performing the tasks or sub-tasks required by the system. One of these decoders will produce a textual summary of the scene.

Potential research directions, aiming to go beyond an audio scene description system by assembling existing bricks, can be discussed and refined with the candidate.

Main activities

- Bibliography, development and evaluation of deep learning systems ;
- Definition of a new task, definition of a corpus and evaluation protocol ;
- Work on the alignment between self-supervised representations of the speech signal and large language models ;
- Weakly supervised system training ;
- System evaluation.

Skills

Master level in computer science, mathematics or phonetics.

Strong interest in applied research.

Written and spoken English

Signal processing

Machine learning and deep learning

Experience with deep learning toolkits such as pytorch or keras

Speech processing experience, knowledge of open source toolkits such as kaldi or speechbrain.

References

[1] Aafaq, N., Mian, A., Liu, W., Gilani, S. Z., & Shah, M. (2019). Video description: A survey of methods, datasets, and evaluation metrics. *ACM Computing Surveys (CSUR)*, *52*(6), 1-37.

[2] Hori, Chiori, Huda Alamri, Jue Wang, Gordon Wichern, Takaaki Hori, Anoop Cherian, Tim K. Marks, et al. « End-to-End Audio Visual Scene-Aware Dialog Using Multimodal Attention-Based Video Features ». In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2352□56. Brighton, United Kingdom: IEEE, 2019. https://doi.org/10.1109/ICASSP.2019.8682583.

[3] Zhang, C., & Tian, Y. (2016, December). Automatic video description generation via lstm with joint two-stream encoding. In *2016 23rd International Conference on Pattern Recognition (ICPR)* (pp. 2924-2929). IEEE.

[4] Pratap, Vineel, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, et al. 2023. « Scaling Speech Technology to 1,000+ Languages ». arXiv. http://arxiv.org/abs/2305.13516.

Benefits package

- Subsidized meals,
- Partial reimbursement of public transport costs,
- Leave: 7 weeks of annual leave + 10 extra days off due to RTT (statutory reduction in working hours) + possibility of exceptional leave (sick children, moving home, etc.),
- Possibility of teleworking and flexible organization of working hours,
- Professional equipment available (videoconferencing, loan of computer equipment, etc.),
- Social, cultural and sports events and activities,

Remuneration

- 1st and 2nd year : 2082 € bruts - gross /month
- 3rd year : 2190 € bruts - gross /month

General Information

- **Town/city :** PARIS
- **Inria Center :** Siège
- **Starting date :** 2024-05-01
- **Duration of contract :** 3 years
- **Deadline to apply :** 2024-08-31

Contacts

- **Inria Team :** MIS-DEFENSE (DIRECTION)
- **PhD Supervisor :**
  Maillet Florence / florence.maillet@inria.fr

About Inria

Inria is the French national research institute dedicated to digital science and technology. It employs 2,600 people. Its 200 agile project teams, generally run jointly with academic partners, include more than 3,500 scientists and engineers working to meet the challenges of digital technology, often at the interface with other disciplines. The Institute also employs numerous talents in over forty different professions. 900 research support staff contribute to the preparation and development of scientific and entrepreneurial projects that have a worldwide impact.

Warning : you must enter your e-mail address in order to save your application to Inria. Applications must be submitted online on the Inria website. Processing of applications sent from other channels is not guaranteed.

Instruction to apply

Defence Security :
This position is likely to be situated in a restricted area (ZRR), as defined in Decree No. 2011-1425 relating to the protection of national scientific and technical potential (PPST).Authorisation to enter an area is granted by the director of the unit, following a favourable Ministerial decision, as defined in the decree of 3 July 2012 relating to the PPST. An unfavourable Ministerial decision in respect of a position situated in a ZRR would result in the cancellation of the appointment.

Recruitment Policy :
As part of its diversity policy, all Inria positions are accessible to people with disabilities.