Offer #2024-07309

Doctorant F/H Description automatisée de scènes audio explicable et frugale

The offer description below is in French

Contract type : Fixed-term contract

Level of qualifications required : Graduate degree or equivalent

Fonction : PhD Position

Context

Inria Défense&Sécurité (Inria D&S) a été créé en 2020 pour fédérer les actions d'Inria répondant aux besoins numériques des forces armées et forces de l'intérieur. La thèse sera réalisée au sein de l'équipe de recherche en traitement de l'audio de Inria D&S, sous la direction de Jean-François Bonastre et co-encadrée par Raphaël Duroselle.

La description automatisée de scènes audio consiste à présenter aux opérateurs un condensé des informations présentes dans la scène en question, sous la forme d'un texte augmenté. Ce condensé permet de faire ressortir de façon synthétique et visuelle les informations les plus importantes, tout en structurant efficacement l'accès aux informations précises. Pour illustrer ce point, un condensé pourrait prendre la forme suivante : « Dans cet enregistrement d'une durée de cinq minutes, trois locuteurs différents sont présents. Le locuteur A correspond à une identité connue dans la base de données et s'exprime en Français avec un fort accent du Monawa, les locuteurs B et C sont inconnus dans la base de données et s'expriment en Français dans leurs interactions avec A et dans une langue non identifiée lorsqu'ils parlent ensemble. Les voix de B et C présentent de fortes similitudes avec les locuteurs de la région du Quabar oriental. Le thème général de l'enregistrement concerne un transfert de marchandises entre les villes de Orienta et de Flagrance. La date du 8 Juillet 2023 est citée à trois reprises ». En cliquant sur A, l'opérateur disposera des informations sur A et sur les détails de l'identification vocale réalisée. L'accès aux segments temporels pendant lesquels A a parlé et à la transcription de ceux-ci sera direct. Dans cette transcription, les noms de personnes, de lieux ou les dates (les entités nommées) seront mises en évidence.

Assignment

Objectif

La thèse vise à proposer un cadre général pour le traitement des enregistrements audio dans le cadre du renseignement. Elle consiste à définir une application de haut niveau adaptée aux besoins des utilisateurs finaux promouvant la présentation d'un enregistrement sous la forme d'un rapport synthétique pour mettre en évidence les points saillants.

Approche

L'approche visée s'inspire à la fois de la description textuelle de scènes vidéo [1] et sur les systèmes de dialogue reposant sur des scènes audio-visuelle [2]. Le système reposera sur l'extraction de représentations du signal de parole à différentes échelles (trame, segment de parole ou événement sonore, enregistrement complet), éventuellement dédiées à des tâches différentes. Les représentations, utiles aux différentes briques technologiques du système seront des embeddings extraits de réseaux de neurones profonds, génériques [3] ou dédiés à chaque tâche. La fusion entre les différents niveaux d'information pourra être réalisée avec une architecture s'inspirant du schéma « Encodeur-Decodeur » multi-stream [4], avec plusieurs encodeurs produisant des séquences de représentations et un ou plusieurs décodeurs réalisant les tâches ou sous-tâches nécessaires au système. Un de ces décodeurs produira un descriptif textuel de la scène.

Des directions de recherche potentielles, visant à dépasser un système de description de scènes audio par assemblage de briques existantes, pourront être discutées et affinées avec le candidat.

Main activities

- Etat de l'art, constitution d'un système de description de scènes audio par assemblage des outils existants ;
- Définition de la tâche, élaboration d'un corpus et d'un protocole d'évaluation ;
- Travail sur l'alignement entre des représentations auto-supervisées du signal de parole et des grands modèles de langage ;
- Entraînement faiblement supervisé du système ;
- Evaluation des systèmes et confiance dans les prédictions.

Skills

- Master 2 ou diplôme d'école d'ingénieur en informatique, mathématiques appliquées ou phonétique,
- Intérêt marqué pour la recherche appliquée,
- Maîtrise de l'anglais parlé et écrit,
- Connaissances en traitement du signal,
- Connaissances en apprentissage automatique de manière générale et dans les approches neuronales (deep learning) en particulier,
- Connaissance pratique d'outils comme Pytorch, Keras ou Scikit-learn,
- Expérience en traitement automatique de la parole, dont la connaissance de plateformes open-source comme Kaldi ou Speechbrain.

Références

[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

- Restauration subventionnée,
- Transports publics remboursés partiellement,
- Congés: 7 semaines de congés annuels + 10 jours de RTT (base temps plein) + possibilité d'autorisations d'absence exceptionnelle (ex : enfants malades, déménagement),
- Possibilité de télétravail (2 jours par semaine) et aménagement du temps de travail,
- Équipements professionnels à disposition (visioconférence, prêts de matériels informatiques, etc.),
- Prestations sociales, culturelles et sportives (Association de gestion des œuvres sociales d'Inria),
- Accès à la formation professionnelle,

Remuneration

Année 1 & 2 = 2082 € bruts mensuels

Année 3 = 2190 € bruts mensuels

General Information

- **Town/city :** PARIS
- **Inria Center :** Siège
- **Starting date :** 2024-05-01
- **Duration of contract :** 3 years
- **Deadline to apply :** 2024-09-01

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

Nous vous remercions d'adresser un CV accompagné d'une lettre de motivation.

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.