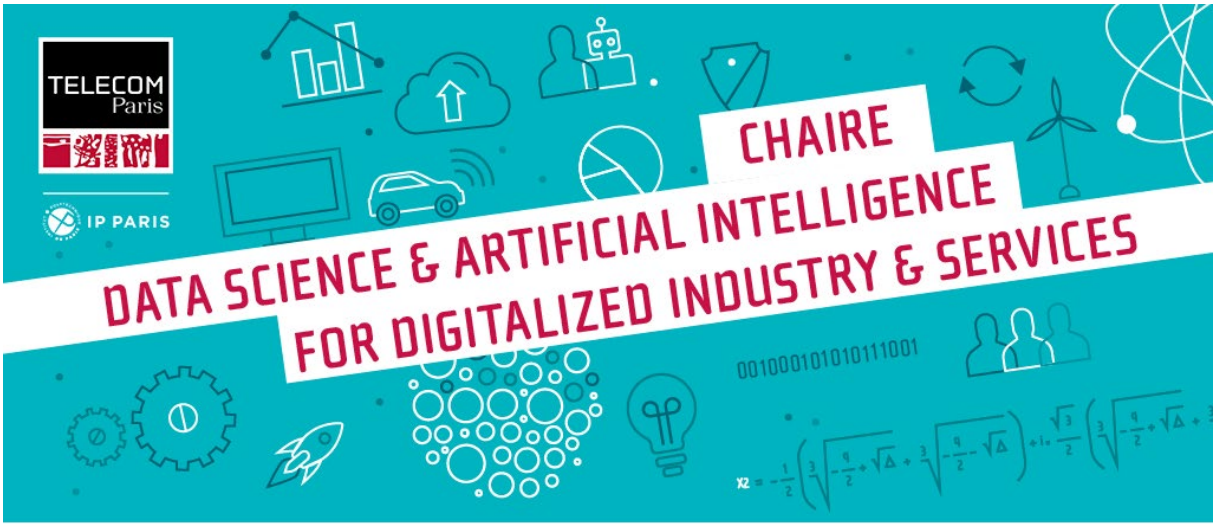

Offre de stage

Fusion multimodale avec modèles d'attention pour la détection de la baisse d'engagement dans les interactions humain-robot

A propos de la chaire

La Chaire Data Science and Artificial Intelligence for Digitalized Industry and Services, portée par Florence d'Alché-Buc, enseignante-chercheuse dans le département Image, Données, Signal de Télécom Paris, la chaire DSAIDIS réunit cinq partenaires industriels : Airbus Defence & Space, Engie, Idemia, Safran et Valeo. Son objectif général est de développer, en liaison étroite avec les partenaires, une formation et une recherche de niveau international.

Ses quatre principaux axes de recherche sont :

1. Analyse et prévision de séries temporelles (Predictive Analytics on Time Series) ;

2. Exploitation de données hétérogènes, massives et partiellement étiquetées (Exploiting Large Scale and Heterogeneous, Partially Labelled Data) ;

3. Apprentissage pour une prise de décision robuste et fiable (Learning for Trusted and Robust Decision) ;

4. Apprentissage dans un environnement dynamique (Learning through Interactions with a Changing Environment).

En savoir plus sur la chaire : **www.telecom-paris.fr/dsaidis**

Description du stage

Encadrement

Chloé Clavel , Asma Atamna

Lieu et dates du stage

Adresse : Télécom Paris, 19 Place Marguerite Perey, 91120 Palaiseau

Date de début du stage : Premier trimestre 2020

Équipe(s) d'accueil de la thèse

Département IDS, équipe Signal, Statistiques et Apprentissage (S2A)

Mots clés

Interaction humain-robot, analyse des sentiments, fusion multimodale, modèles d'attention, réseaux de neurones récurrents, social computing

Sujet détaillé

Le social computing et l'interaction humain-robot sont des domaines de l'informatiques dont l'un des enjeux est la conception d'agents virtuels capables de démontrer des comportements sociaux. Les robots d'assistance sociale (socially assistive robots - SAR), par exemple, sont conçus pour assister les humains dans différentes tâches telles que le coaching, l'aide thérapeutique ou l'orientation des voyageurs dans les gares et aéroports. En tant que tels, les SAR doivent être en mesure d'afficher des comportements sociaux afin que l'utilisateur se sente engagé dans l'interaction.

L'efficacité d'un agent virtuel dépend de sa capacité à interpréter correctement les émotions de l'utilisateur en analysant les données séquentielles—et souvent multimodales (audio, vidéo, texte)—de l'interaction. Une question particulièrement importante dans ce contexte est comment combiner—ou fusionner—les données des différentes modalités afin de modéliser au mieux les dépendances entre ces dernières.

La fusion multimodale [Atrey10] se fait généralement selon l'une des deux approches classiques suivantes : (i) la early fusion qui consiste à concaténer les différentes entrées multimodales avant de les passer au modèle et (ii) la late fusion où un modèle est appris pour chaque modalité et les sorties des différents modèles sont ensuite combinées pour obtenir la décision finale. Récemment, cependant, différentes approches utilisant des modèles d'attention pour effectuer la fusion multimodale ont été proposées afin de pallier les limites des early et late fusions [Zadeh18a,Zadeh18b,Gu18,Tsai19]. Ces approches utilisent l'attention pour modéliser les dynamiques au sein de chaque modalité (dynamiques intra-modalité), ainsi que les dynamiques entre les différentes modalités (dynamiques inter-modalité). De manière générale, un modèle d'attention [Bahdanau15] apprend des scores (ou poids) d'attention pour chaque élément d'une séquence donnée en entrée. Ces scores sont ensuite utilisés lors de la prédiction pour "se concentrer" sur les parties les plus importantes de la séquence.

Ce stage viendra renforcer l'axe 4 de la chaire DSAIDIS dans le cadre des interactions humain-robot. Il porte sur le développement d'approches de fusion multimodale basées sur des modèles d'attention afin d'améliorer la détection de la baisse d'engagement dans le cadre d'interactions humain-robot. Partant des travaux de [Ben-Youssef19], le stagiaire développera des architectures neuronales récurrentes qui utiliseront des modèles d'attention pour effectuer la fusion multimodale en la présence de deux modalités : audio et vidéo. Plus précisément, le stagiaire sera amené à :

- Etudier la littérature récente sur la fusion multimodale, les modèles d'attention et la reconnaissance des émotions,
- Manipuler les données utilisateur [Ben-Youssef17] afin de séparer les modalités audio et vidéo,
- Implémenter des réseaux de neurones neurones récurrents basés sur des unités LSTM ou GRU pour la détection de la baisse d'engagement dans les séquences de données audio et vidéo,
- Implémenter et évaluer des modèles d'attention pour la fusion multimodale dans le cas de données audio et vidéo.

Profil du candidat

- Etudiant titulaire d'un master 2 recherche
- Apprentissage statistique / reconnaissance des formes
- Traitement de la parole, traitement du langage naturel
- Bon niveau en programmation (Java, C/C++, Python)
- Bon niveau d'anglais

Candidatures

A envoyer en un seul pdf à chloe.clavel@telecom-paristech.fr, asma.atamna@telecom-paristech.fr.

- Curriculum Vitae
- Lettre de motivation personnalisée expliquant l'intérêt du candidat sur le sujett
- Relevés de notes des années précédentes
- Contact d'une personne de référence

Les candidatures incomplètes ne seront pas examinées.

Références

[Atrey10] Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, Mohan S. Kankanhalli. Multimodal Fusion for Multimedia Analysis: A Survey. Multimedia Systems 2010.

[Bahdanau15] Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. ICLR 2015.

[Ben-Youssef17] Atef Ben-Youssef, Chloé Clavel, Slim Essid, Miriam Bilac, Marine Chamoux, Angelica Lim. UE-HRI: A New Dataset for the Study of User Engagement in Spontaneous Human-Robot Interactions. ICMI 2017.

[Ben-Youssef19] Atef Ben-Youssef, Giovanna Varni, Slim Essid, Chloé Clavel. On-the-Fly Detection of User Engagement Decrease in Spontaneous Huma-Robot Interaction Using Recurrent and Deep Neural Networks. Int. J. of Soc. Robotics 2019.

[Gu2018] Yue Gu, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li, and Ivan Marsic. Mutlimodal Affective Analysis Using Hierarchical Attention Strategy with Word-Level

Alignment. ACL 2018.

[Tsai19] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal Transformer for Unaligned Multimodal Language Sequences. ACL 2019.

[Zadeh18a] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory Fusion Network for Multi-view Sequencial Learning. AAAI 2018.

[Zadeh18b] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. Multi-attention Recurrent Network for Human Communication Comprehension. AAAI 2018.