Titre: Interprétabilité des modèles pré-entrainés pour le traitement automatique de la parole

Encadrement: Nicolas Dugué, Anthony Larcher (direction) et Marie Tahon (co-direction)

Laboratoire: Laboratoire d'Informatique de l'Université du Mans (LIUM), site du Mans

Début de la thèse : dès que possible

Avec l'avènement des réseaux de neurones, les travaux en apprentissage automatique se sont éloignés des méthodes traditionnelles qui nécessitaient une grande expertise des données, couramment appelées *feature engineering* [1]. En effet, outre leurs performances, les réseaux de neurones permettent de formuler mathématiquement le problème à résoudre, et d'appliquer des algorithmes d'optimisation génériques afin d'en obtenir une solution. La solution émerge donc naturellement des données et du critère à optimiser : l'expertise humaine, auparavant importante quant aux données se situe maintenant essentiellement sur la formulation du problème. Les approches neuronales permettent ainsi de traiter différents types de données avec des chaînes de traitement génériques dont la première brique est souvent l'apprentissage de représentation vectorielles pour les données.

Dans ce cadre, la thèse s'inscrit dans un champ de recherche qui cherche à reprendre contact avec l'intelligence artificielle telle qu'elle existait avec le *feature engineering*, celui de l'interprétabilité. Il s'agit dans ce champ de recherche de comprendre et d'expliquer les modèles neuronaux et leurs performances, en reconnectant les résultats aux données et à leurs attributs interprétables humainement. En particulier, le travail de thèse concerne l'apprentissage de représentations vectorielles, à la racine de tous les chaînes de traitement. Dans le cadre de l'apprentissage de représentation pour les données textuelles, la recherche a fait émerger des pistes intéressantes : apprendre dans des espaces de plus grande taille [2,3], forcer la parcimonie [4], analyser les représentations internes [5,6]. L'objet texte est facilement interprétable par l'être humain, et est naturellement discret. La nature même du signal audio rend le travail d'interprétation plus difficile. En effet, c'est un signal continu de grande dimension où des informations de différentes natures se superposent à des échelles de temps très différentes: les descripteurs acoustiques de bas-niveau (trame), la prononciation (phone), la linguistique (mot), l'expressivité (phrase). De très récents travaux, ont permis de développer des visualisations du signal permettant d'interpréter certains aspects des signaux audio, la plupart utilisant des modèles locaux comme SHAP [7,8]. Les visualisations sont intéressantes, mais elles nécessitent une forte compréhension de l'outil et sont difficiles à exploiter à grande échelle. Dans cette thèse, nous souhaitons explorer les représentations vectorielles utilisées pour le traitement de la parole (WavLM, X-vector, etc) pour être capable de les interpréter avec des descripteurs experts connus sans avoir à ré-entrainer des modèles consommateurs de ressources tels que WavLM. L'utilisation de la synthèse de parole à partir de ces représentations permettra d'évaluer à quel point les interprétations obtenues automatiquement sont réellement interprétables par des humains.

Afin d'avancer dans la construction de systèmes interprétables exploitant du signal audio, forts de notre expérience dans le cadre textuel, nous souhaitons initier des travaux sur l'audio avec ce sujet de thèse de doctorat.

Le sujet se découperait en plusieurs axes :

1) explorer les espaces de plongements extraits sur différentes fenêtres temporelles (de la trame à la phrase) afin de mettre au jour des dimensions latentes interprétables (attributs prosodiques, phonétiques, locuteur et linguistique), en nous inspirant de [5], permettant ainsi de reconnecter les espaces abstraits à des attributs du signal de parole humainement compréhensibles, voire de construire une approche (par exemple à partir de [9]) permettant d'obtenir un mapping bijectif entre la représentation vectorielle et les descripteurs experts ; la mise à jour des dimensions latentes interprétables passera par la création de jeux de données de parole naturelle ou synthétique adaptés aux attributs recherchés ;

2) évaluer la robustesse de ces approches quant à l'apprentissage, et en particulier, leur capacité à exploiter de façon régulière ou non les descripteurs exploités dans le premier axe, il faudra pour cela mettre en place des méthodologies d'évaluation quantitative et perceptive ;

3) décrire de nouveaux modèles d'apprentissage de plongements de trame interprétables par construction en forçant une correspondance avec les attributs détectés lors du premier axe ;

4) appliquer ce travail à la sécurité, l'anonymisation des données, et/ou la suppression de biais : des représentations interprétables sont plus facilement manipulables pour cacher des éléments non désirables dans la représentation vectorielle à partir des descripteurs experts.

[1] Cardon, Dominique, et al. "Neurons spike back." Réseaux 211.5 (2018): 173-220.

[2] Subramanian, Anant, et al. "Spine: Sparse interpretable neural embeddings." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 32. No. 1. 2018.

[3] Prouteau, Thibault, et al. "SINr: Fast Computing of Sparse Interpretable Node Representations is not a Sin!." International Symposium on Intelligent Data Analysis. 2021.

[4] Murphy, Brian, Partha Talukdar, and Tom Mitchell. "Learning effective and interpretable semantic models using non-negative sparse embedding." Proceedings of COLING 2012. 2012.

[5] Bolukbasi, Tolga, et al. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings." Advances in neural information processing systems 29 (2016).

[6] Clark, Kevin, et al. "What does bert look at? an analysis of bert's attention." arXiv preprint arXiv:1906.04341 (2019).

[7] W. Ge, J. Patino, M. Todisco and N. Evans, "Explaining Deep Learning Models for Spoofing and Deepfake Detection with Shapley Additive Explanations," ICASSP 2022, pp. 6387-6391 (2022).

[8] Sivasankaran, E. Vincent and D. Fohr, "Explaining deep learning models for speech enhancement," in Proc. Interspeech, pp. 696–700 (2021).

[9] Paul-Gauthier Noé, Andreas Nautsch, Driss Matrouf, Pierre-Michel Bousquet, Jean-François Bonastre. A bridge between features and evidence for binary attribute-driven perfect privacy. ICASSP 2022, May 2022, Singapore, Singapore

Profil du candidat : Le candidat devra être motivé pour travailler sur le langage écrit et parlé, et montrer un intérêt pour la synthèse de parole. Il devra avoir Master en Informatique, une expérience en machine learning sera appréciée.
Pour candidater : Envoyer CV + lettre de motivation avant le 15 octobre 2023 à :

Nicolas.dugue@univ-lemans.fr; marie.tahon@univ-lemans.fr