Expliquer les prédictions des réseaux de neurones par l'exploration de l'espace de représentation et de la frontière de décision à l'aide d'EBBE-Text

Alexis Delaforge* Jérôme Azé* Arnaud Sallaberry*,**
Maximilien Servajean*,** Sandra Bringay*,** Caroline Mollevi***,****

*LIRMM, Université de Montpellier, CNRS
CC477 - 161 rue Ada, 4095 Montpellier Cedex 5, France
prenom.nom@lirmm.fr
http://www.lirmm.fr/
**Groupe AMIS, Université Paul-Valéry Montpellier 3
Route de Mende, 34199 Montpellier Cedex 5, France
***Institut du Cancer Montpellier (ICM)
208 Avenue des Apothicaires, Parc Euromédecine, 34298 Montpellier Cedex 5, France
caroline.mollevi@icm.unicancer.fr,
https://www.icm.unicancer.fr/fr
****Institut Desbrest d'Epidémiologie et de Santé Publique,
UMR Inserm - Université de Montpellier, Montpellier, France

Résumé. En classification automatique de textes, de nombreux travaux récents portent sur l'interprétation des réseaux de neurones par la production d'explications associées aux prédictions. Dans ce contexte, EBBE-Text offre une visualisation interactive de la frontière de décision, du positionnement des textes vis-à-vis de celle-ci (et donc de la certitude d'un réseau en ses prédictions), des chemins menant d'un texte à la frontière de décision, des informations concernant la proximité entre les textes, tout cela au sein de différentes localités dans l'espace de représentation des textes. Ces informations permettent d'intuiter comment le réseau de neurones de classification fonctionne et ainsi aider à son interprétabilité. Notre méthode crée des données sur la frontière de décision puis utilise des ensembles flous simplicials pour créer un graphe avant d'aligner linéairement les données créées sur la frontière de décision. Enfin, un processus itératif place les données d'entrée autour des arrangements linéaires des données de la frontière.

1 Introduction

Récemment, les réseaux de neurones ont connu un vif succès dans les tâches de Traitement Automatique du Langage (TAL) comme la traduction, la reconnaissance d'entités nommées ou encore l'analyse de sentiments. L'utilisation des techniques d'apprentissage profond soulèvent des questions sur l'interprétabilité et l'explicabilité de ces réseaux (Lipton, 2018). Il est, d'ailleurs, primordial de s'intéresser à ces questions, d'autant plus que le parlement européen