EBBE-Text : Visualisation de la frontière de décision des réseaux de neurones en classification automatique de textes

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 des prédictions. L'approche originale présentée dans cet article consiste à visualiser la frontière de décision et le positionnement des données vis-à-vis de celle-ci offrant ainsi une nouvelle approche de l'explication. Notre méthode calcule tout d'abord un espace de représentation des phrases, puis en exploite la structure linéaire afin de visualiser la répartition et le regroupement des données autour de la frontière de décision. Le principal apport de notre méthode est le processus de visualisation de la frontière de décision, permettant d'explorer la distance à la frontière de décision (et donc la certitude d'un réseau en ses prédictions) mais aussi les chemins menant à celle-ci ou encore la proximité entre phrases.

1 Introduction

Récemment, les réseaux de neurones ont connu du succès dans les tâches de Traitement Automatique du Langage (TAL) (Young et al., 2018) comme la traduction (Cho et al., 2014; Bahdanau et al., 2015), la reconnaissance d'entités nommées (Collobert et al., 2011) ou encore l'analyse de sentiments (Socher et al., 2013). L'utilisation des techniques d'apprentissage profond soulève des questions sur l'interprétabilité, l'explicabilité, la confiance et la transparence de ces réseaux (Lipton, 2018). Il est, d'ailleurs, primordial de s'intéresser à ceux-ci, d'autant plus que le parlement européen (Goodman et Flaxman, 2016) a fixé des règles parmi les plus strictes au monde concernant l'interprétabilité de ces réseaux de neurones.