N° d'ordre : D -

THESE

présentée

devant l'Institut National des Sciences Appliquées de Rennes

en vue de l'obtention du

DOCTORAT

spécialité : Informatique

par M. Duc Hau NGUYEN

Intitulé : Making AI Understandable for Humans: the Plausibility of Attention-Based Explanations in Natural Language Processing

Directeur de Thèse : Pascale Sébillot & Guillaume Gravier

Date, heure et lieu de soutenance : le 11 octobre 2024 à 13h45 à l'IRISA
Salle Hélvior

Membres du jury (nom, prénom, titre et établissement de rattachement, fonction)

- CERISARA Christophe, Chargé de recherche, CNRS, HDR, Loria, rapporteur
- MORIN Emmanuel, Professeur des universités, Nantes Université, LS2N, examinateur
- TANNIER Xavier, Professeur des universités, Sorbonne Université, Limics, rapporteur
- SÉBILLOT Pascale, Professeur des universités, INSA Rennes, IRISA, directrice de thèse
- GRAVIER Guillaume, Directeur de recherche, CNRS, IRISA, co-directeur de thèse

RESUME DE LA THESE

Originally designed to explain deep learning models, explainable artificial intelligence (XAI) should also provide accessible explanations to assist end users in their domains of expertise. This thesis explores, within the context of natural language processing, the potential of using the attention mechanism—a crucial component of recent neural architectures—to provide users with plausible and easily understandable explanations for the decisions made by an algorithm. We first show that attention tends to be distributed across all the words in an utterance, making the explanations poorly plausible. Inspired by the specificities of human annotations seeking to explain decisions, we first propose a morphosyntactic filtering strategy that concentrates attention on the sole nouns, verbs, and adjectives present in the utterances. To obtain a more generalizable solution that does not depend on specific dataset characteristics, we further introduce three regularization strategies during the training phase, based either on an entropy criterion or on supervised or unsupervised approaches. Finally, our thesis reveals that deep architectures allowing for contextualization of word representations in utterances tend to penalize the explainability of attention due to their convergence within the representation space. By minimizing the influence of contextualization on words, the attention mechanism can more effectively approach explanations produced by humans.