Vers une IA explicable pour les systèmes critiques

PICARD, Agustin Martin, agustinmartin.picard@scalian.com, Scalian, Toulouse

Résumé : *Au cours de l'année 2020, une équipe de chercheurs et data scientists industriels du projet DEEL s'est attaqué au problème de l'explicabilité des modèles d'intelligence artificielle dans le cadre des applications critiques. Nous présentons quelques résultats que nous avons eus grâce à cette composition mixte.*

Mots clés : *Explainability, Interpretability, Deep Learning, Computer Vision*

1. Introduction

Le projet DEEL (DEpendable and Explainable Learning) est un projet franco-canadien regroupant 17 partenaires et qui compte avec le soutien d'ANITI – un des 4 instituts créés par l'initiative 3IA – et un budget de 30M€. Son objectif est de promouvoir les échanges entre les groupes industriels et les laboratoires de recherche pour stimuler les avancées techniques sur les problématiques en commun. Dans ce contexte-là, nous avons une équipe dédiée à répondre à la problématique de la compréhension du fonctionnement des modèles d'IA.

2. Méthodologie

L'une des problématiques les plus récurrentes chez les industriels est celle de l'interprétabilité des modèles d'intelligence artificielle de type boîte noire. En effet, il existe tout un processus de certification de sûreté des systèmes critiques pour garantir qu'ils ne mettent pas en danger la vie des utilisateurs, et donc la capacité de comprendre le fonctionnement du modèle est essentielle. De ce fait, une équipe composée de data scientists industriels et chercheurs s'est mise à développer et mettre à l'épreuve toute une batterie de techniques différentes.

Commençant par une étude bibliographique de l'état de l'art, nous avons identifié les différentes familles de méthodes, et avec ces informations nous avons déterminé celles qui pourraient répondre à ce besoin d'interprétabilité rigoureuse. Pour les techniques les plus prometteuses, nous évaluons leur pertinence en les appliquant aux cas d'usage industriels DEEL, et si besoin, nous implémentons d'éventuelles améliorations.

En particulier, nous avons commencé par analyser une méthode issue du domaine de l'analyse de sensitivité qui est capable de générer des explications globales pour des données tabulaires et des images.

Dans un second temps, une de nos équipes a étudié la pertinence des méthodes d'attribution pour des modèles de réseaux de neurones convolutionnels pour des données de type image, et la mesure dans laquelle elles sont susceptibles de générer des explications qui ne dépendent pas du modèle, et donc, qui n'éclairent pas vraiment ses prédictions.

En parallèle, une deuxième équipe a considéré les « Variational AutoEncoders » (VAEs), des modèles de réseaux de neurones capables d'apprendre des représentations de façon non-supervisée, pour déterminer l'intérêt des représentations désenchevêtrées pour l'interprétabilité des réseaux de neurones dans des applications de vision par ordinateur.

Enfin, nous avons développé une métrique d'explicalibilité qui nous permet de déterminer dans quelle mesure une méthode est capable de fournir des explications consistantes d'une part, et la capacité du modèle lui-même à apprendre des stratégies qui généralisent sur les différents sous-groupes d'une classe de l'autre.

3. Originalité / perspective

La recherche sur l'explicabilité des modèles d'IA s'est focalisé jusque-là sur des problématiques théoriques qui ne sont pas forcément facilement transférables aux cas d'usages industriels. Nous étudions l'applicabilité de différentes techniques d'explicabilité dans le cadre des systèmes critiques où la capacité de comprendre le modèle joue un rôle crucial dans le processus de certification de sûreté des systèmes.

Site de DEEL : https://www.deel.ai/