IA Explicable et qualité des données/modèles

Offre de stage de master 2ème année ou stage de fin d'étude d'ingénieur en informatique – Année 2025

Localisation :
Laboratoire GREYC, CNRS UMR 6072, Université de Caen Normandie, 14000, Caen

Contexte scientifique

Ce stage de master s'inscrit dans le cadre du projet PANDORA financé par l'ANR (Agence Nationale de la Recherche), projet qui démarrera en février 2025. PANDORA se situe dans le contexte de l'intelligence artificielle explicable (XAI), en particulier dans le domaine des réseaux de neurones sur graphes (GNN). En se focalisant sur le fonctionnement interne des GNNs, les objectifs du projet sont les suivants :

- caractériser, comprendre et expliquer de manière claire le fonctionnement interne des GNN en utilisant des techniques d'extraction de motifs ;
- découvrir des motifs d'activation neuronale statistiquement significatifs, appelés « règles d'activation », pour déterminer comment les réseaux encodent les concepts [7, 8] ;
- traduire ces règles d'activation en motifs de graphes interprétables par un utilisateur ;
- utiliser ces connaissances pour améliorer les GNN en identifiant les biais d'apprentissage, en générant des données supplémentaires et en construisant des systèmes d'explication.

Ce stage de recherche porte sur le dernier point. Plus précisément, nous souhaitons développer de nouvelles méthodes permettant d'améliorer l'apprentissage des modèles sur graphes en s'appuyant sur l'analyse du fonctionnement interne de ces modèles via, par exemple, des règles d'activation exprimées dans l'espace latent. Il s'agira ainsi d'analyser les frontières de décisions, de caractériser les erreurs du modèle étudié dans l'espace des données ou dans leurs représentations latentes afin de proposer des solutions correctives. Cette approche peut se décomposer en sous-problèmes :

Caractérisation de données et identification de biais. La caractérisation des données d'apprentissage peut permettre d'identifier les instances sur lesquelles le modèle fait des erreurs mais également détecter si les données ne sont pas à l'origine de biais dans l'apprentissage. Une piste de travail est d'étudier la complexité des règles d'activation et de les comparer aux connaissances du domaines.

Génération ciblée de données supplémentaires. Une fois les limites du modèle identifiées, nous souhaitons définir de façon automatique des "patchs correctifs" afin d'améliorer la robustesse du modèle. Un axe de travail privilégié sera la génération

de données supplémentaires ciblées pour permettre au modèle de mieux séparer les données selon la classe étudiée dans la représentation construite.

La deuxième problématique pose des questions de recherche relativement complexe et elle sera traitée pendant la thèse qui se déroulera à la suite de ce stage dans le contexte de PANDORA. Le travail, dans ce stage, se focalisera sur la première problématique, c.-à-d. la caractérisation de données. Les résultats obtenus dans ce stage serviront de pierre angulaire à la recherche qui sera développée dans la thèse.

Problématique

En apprentissage automatique, on ne dispose pas toujours d'ensembles de données d'apprentissage suffisamment représentatif du monde réel (par exemple, les expériences chimiques/biologiques se focalisent souvent seulement sur certaines molécules bien explorées ou certaines cibles thérapeutiques). Comment détecter qu'un ensemble de données d'apprentissage est insuffisant ? Deux propositions non exhaustives pour cela :

— des parties de l'espace de données possibles ne sont pas représentées (p.ex. certaines combinaisons de noeuds/arêtes ne peuvent pas être trouvées).

— le modèle appris est peu fiable dans certains sous-espaces des données (la fiabilité d'un modèle supervisé peut être étudiée, par exemple, en regardant l'importance des instances dans la construction des frontières de décision).

La littérature contient des méthodes pour caractériser les données d'une manière indépendante du modèle [5] et des méthodes pour caractériser le comportement d'un modèle en fonction des composantes des graphes individuels considérés [9, 2, 6, 3, 4, 1]. Cependant, il n'existe pas d'approche qui établit le lien entre les données et la performance d'un modèle spécifique. Ce point est la question de recherche à traiter dans ce stage.

Le travail mené dans ce stage s'appuiera sur des données moléculaires résultant d'expériences biochimiques issues de notre collaboration avec le laboratoire CERMN (Centre d'Études et de Recherche sur le Médicament de Normandie), Université de Caen Normandie.

Objectifs

Ce stage comporte deux objectifs. Dans un premier temps, nous souhaitons concevoir une (ou plusieurs) approches pour utiliser les explications du comportement des GNNs afin d'identifier les instances pertinentes du jeu d'entraînement utilisé. Puis, nous voulons caractériser à un niveau global les jeux de données de graphes d'une manière similaire à celle déjà utilisée pour des jeux de données vectorielles.

Étapes et livrables

1. Effectuer une synthèse bibliographique de méthodes d'explication du comportement des modèles GNN [9, 2, 8, 7, 6, 3, 4, 1]. Le but de cette étude est d'établir dans quel sens les différentes méthodes identifient certains aspects des données utilisées pour entraîner le modèle.
2. Concevoir et implémenter des approches pour identifier les instances (les graphes) impliquées par les descripteurs/règles explicatifs. Il n'est pas certain que de telles approches soient trouvées pour tous les descripteurs, ce qui conduira alors à une

sélection de descripteurs. La mise en évidence des instances et graphes liés aux descripteurs/règles explicatifs permettra aussi de déterminer comment les descripteurs caractérisent différents sous-ensembles de données.

3. Développer un formalisme afin d'étendre des concepts définis pour des données vectorielles (densité, frontières de décision, distribution de valeurs) à des données de type graphe. Ce formalisme, en combinaison avec les résultats de l'étape 2, permettra de déterminer où les instances d'apprentissage manquent dans un jeu de données d'entraînement et ainsi où il est utile de générer des données synthétiques.

Mots clés : Apprentissage statistique, réseaux de neurones sur graphes, IA explicable, fouille de données.

Période : Stage de 5 à 6 mois à effectuer entre le 1er mars et le 31 août 2025.

Gratification : Selon les règles en vigueur (environ 650€ par mois).

Équipe d'encadrement :

— Bruno Crémilleux (GREYC – Université de Caen Normandie).
— Marc Plantevit (LRE – EPITA)
— Albrecht Zimmermann (GREYC – Université de Caen Normandie).

Pour candidater :

Envoyer les documents suivants (exclusivement au format pdf) à `bruno.cremilleux@unicaen.fr`, `marc.plantevit@epita.fr` et `albrecht.zimmermann@unicaen.fr` :

— lettre de motivation expliquant vos qualifications, expériences et motivation pour ce sujet ;
— curriculum vitae ;
— relevé de notes (si possible avec classement) de licence 3, de 1ère année de master et les notes de 2ème année de master disponibles ou équivalent pour les écoles d'ingénieurs ;
— si possible, noms de personnes (enseignants ou autre personne) pouvant fournir des informations sur vos compétences et votre travail ;
— un lien vers des dépôts de projets personnels (par exemple GitHub) ;
— toute autre information que vous estimerez utile.

Références

[1] C. Abrate, G. Preti, and F. Bonchi. Counterfactual explanations for graph classification through the lenses of density. In *World Conference on Explainable Artificial Intelligence*, pages 324–348. Springer, 2023.

[2] A. Duval and F. D. Malliaros. Graphsvx : Shapley value explanations for graph neural networks. In *Machine Learning and Knowledge Discovery in Databases. Research Track : European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part II 21*, pages 302–318. Springer, 2021.

[3] Q. Huang, M. Yamada, Y. Tian, D. Singh, and Y. Chang. Graphlime : Local interpretable model explanations for graph neural networks. *IEEE Transactions on Knowledge and Data Engineering*, 35(7) :6968–6972, 2022.

[4] A. Mastropietro, G. Pasculli, C. Feldmann, R. Rodríguez-Pérez, and J. Bajorath. Edgeshaper : Bond-centric shapley value-based explanation method for graph neural networks. *Iscience*, 25(10), 2022.

[5] M. A. Munoz, L. Villanova, D. Baatar, and K. Smith-Miles. Instance spaces for machine learning classification. *Machine Learning*, 107(1) :109–147, 2018.

[6] A. Perotti, P. Bajardi, F. Bonchi, and A. Panisson. Graphshap : Explaining identity-aware graph classifiers through the language of motifs. *arXiv preprint arXiv :2202.08815*, 2022.

[7] L. Veyrin-Forrer, A. Kamal, S. Duffner, M. Plantevit, and C. Robardet. In pursuit of the hidden features of gnn's internal representations. *Data & Knowledge Engineering*, 142 :102097, 2022.

[8] L. Veyrin-Forrer, A. Kamal, S. Duffner, M. Plantevit, and C. Robardet. On gnn explainability with activation rules. *Data Mining and Knowledge Discovery*, pages 1–35, 2022.

[9] H. Yuan, H. Yu, J. Wang, K. Li, and S. Ji. On explainability of graph neural networks via subgraph explorations. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12241–12252. PMLR, 18–24 Jul 2021.