

Decimation de graphes pour les réseaux profonds sur graphes

Une bourse de thèse au laboratoire GREYC (Caen, France)

Mots clés : Deep learning, Pooling, Graph Decimation, Graph Neural Networks.

Contexte La plupart des objets de notre vie courante sont basés sur des objets discrets avec des relations séquentielles (chaînes de caractères) ou plus complexes (graphes). On peut évoquer les relations entre les personnes dans des graphes sociaux, les liens entre les atomes d'une molécule ou la distance topographique entre les capteurs de vitesse dans le cadre de la prédiction du trafic routier, pour n'en citer que quelques-uns. La prédiction des propriétés de tels objets relève de la reconnaissance structurelle de formes. Pendant des décennies, ce domaine de recherche a été limité par des métriques coûteuses (par exemple, basées sur l'isomorphisme de sous-graphes) ou peu efficaces, généralement combinées à des algorithmes d'apprentissage limités (principalement l'algorithme des k plus proches voisins). Une première percée importante a été réalisée par l'introduction de méthodes à noyaux appliquées aux objets discrets tels que les chaînes de caractères ou les graphes. En plus de fournir des métriques efficaces sur ces objets discrets, ces derniers constituent une porte d'entrée vers de nombreuses méthodes d'apprentissage automatique. Ainsi, ils réduisent l'écart entre les techniques de reconnaissance des formes structurelles et statistiques. Une deuxième avancée dans ce domaine a été fournie par l'introduction des réseaux neuronaux sur graphes (GNN). Comme les noyaux sur graphes, ces réseaux fournissent une connexion solide entre les graphes et les techniques d'apprentissage. De plus, comme d'autres techniques d'apprentissage profond, les GNN évitent de concevoir manuellement une mesure de similarité entre graphes. Les GNN reposent sur deux opérations, à savoir la convolution et la décimation des graphes. Cependant, ces deux opérations présentent encore de graves inconvénients. tout d'abord, le pouvoir expressif des opérations de convolution sur graphes est limitée dans le domaine spectral et correspond généralement à un filtre passe-bas. Deuxièmement, l'opération de décimation du graphe est généralement effectuée par les algorithmes de clustering sur graphes existants, tandis que l'opération équivalente dans les réseaux neuronaux d'image correspond à un sous-échantillonnage, qui offre des garanties en termes de décimation et de connectivité des entités fusionnées.

Ce doctorat se concentrera sur ce dernier problème en étroite collaboration avec d'autres partenaires qui étudient le cadre de la convolution sur graphes.

Sujet de thèse Il convient tout d'abord de distinguer deux concepts : La décimation de graphes, qui consiste à réduire la taille d'un graphe en regroupant des ensembles de sommets connectés, et le pooling de graphes, qui consiste à résumer un graphe connecté par une valeur numérique ou un vecteur.

La thèse sera grossièrement décomposée en trois étapes :

1. **Décimation de graphes :** Le doctorant devra d'abord étudier les techniques de décimation de graphes développées par notre équipe afin de les transposer à une implémentation GPU et au cadre de l'apprentissage profond.

Ces schémas de décimation doivent assurer :

- (a) Un taux de décimation fixe (rapport entre les tailles de deux graphes successifs),
 - (b) Un rayon limité (petit) des sous-graphes regroupés en un seul sommet par le schéma de décimation.
2. **Propriétés spectrales des graphes** : Le doctorant devra étudier la littérature relative aux schémas de décimation préservant les propriétés spectrales des graphes. Il devra ensuite proposer de nouveaux algorithmes combinant les résultats de l'étape précédente avec ces techniques, afin d'assurer la préservation des propriétés spectrales des graphes (notion à affiner) avec un taux de décimation fixe et des tailles bornées de sous-graphes.
3. **Apprentissage de la décimation** : Cette dernière étape est certainement l'une des plus importantes. Les techniques existantes qui apprennent un schéma de décimation fournissent des graphes presque complets, éliminant ainsi la structure du graphe. Le doctorant devra comprendre ces méthodes et les améliorer afin de préserver les propriétés structurelles du résultat en se basant sur les résultats précédents.

Application. La principale application de la thèse sera la conception de nouvelles méthodes facilitant la conception de médicaments en étroite collaboration avec un autre laboratoire spécialisé dans ce domaine. Plus précisément, les applications visées sont la maladie d'Alzheimer et la polypharmacologie. Deux cibles complémentaires seront étudiées pour cette application : 1) améliorer l'activité de prédiction de l'action des molécules (médicaments) sur les ensembles de données fournis par notre partenaire biologique, 2) regrouper automatiquement les atomes en pharmacophores (groupes d'atomes connus pour avoir une activité biologique) grâce à notre schéma de décimation. Notons que notre partenaire biologique a déjà identifié un ensemble de règles artisanales permettant de regrouper les atomes dans des pharmacophores. Notre objectif ici sera de proposer de nouveaux groupements d'atomes grâce à notre schéma de décimation appris.

Profil du candidat Curieux, têtu et autonome le candidat doit avoir un diplôme de master ou d'ingénieur en informatique ou mathématiques appliqués. Une première expérience (cours, stage, projets) en apprentissage et deep learning seraient appréciés. Des compétences complémentaire en théorie des graphes ou un intérêt pour ce domaine seraient un plus.

Noms des contacts

Luc Brun : luc.brun@unicaen.fr (PhD's encadrant)

Benoit Gaüzère : benoit.gauzere@insa-rouen.fr (co encadrant)

Information sur le poste

Localisation : Caen, situé au nord-ouest de la France (2h de Paris).

Salary : au moins 1768 € brut (correspond à 1430€ net). Des décrets ministériels devraient amener à une revalorisation du salaire en 2022.

Limit date to postulate : Octobre 2021

Start of the PhD : Octobre/Novembre 2021.