Sujet de thèse : Apprentissage de représentation de graphes

Graph representation learning

Laboratoire :
— LAMSADE, Université Paris Dauphine, PSL

Direction :
— Florian Yger, Mcf, Université Paris Dauphine (florian.yger@dauphine.fr)
— Virginie Gabrel, Mcf HDR, Université Paris Dauphine (virginie.gabrel@dauphine.fr)

Contexte

Jusqu'à récemment, le domaine de l'apprentissage automatique s'était surtout concentré sur le développement d'approches dédiées aux données euclidiennes. Mais ces progrès sont maintenant graduellement transférés à des données complexes, structurées [7] et plus généralement non-euclidiennes [4]. Les géométries mises en œuvre pour manipuler ces données impliquent des notions d'invariance complexes et souvent coûteuses en temps de calcul. Dans ce contexte, le choix de la représentation de ces données est fondamental. Une représentation bien choisie permettra ainsi d'induire les invariances nécessaires tout en permettant l'application d'algorithmes d'apprentissage classiques.

Dans le cadre de cette thèse, nous nous proposons d'étudier la représentation de graphes [5] dans un cadre supervisé en privilégiant des approches d'apprentissage de métrique [1] et de dictionnaire (comme amorcé dans [3]). Ces pistes seront à mettre en lien avec les approches à noyaux [9, 2] et les réseaux de neurones [12, 7, 10] proposées dans la litérature.

Problématique

Comparer deux graphes revient à trouver les structures équivalentes entre eux et peut se formuler de nombreuses façons (isomorphisme de graphe ou de sous-graphes, appariemment quadratique, chemin d'édition minimal,...). Dans tous les cas, un graphe subissant une permutation de ses sommets ne verra pas sa structure modifiée et devra être considéré comme similaire à sa version non permutée. Toute opération de comparaison de graphe devra donc être invariante aux permutation.
Il existe de nombreuses façons de représenter un graphe (matrices d'adjacence, matrice d'incidence,...) mais rares sont celles qui induisent une invariance aux permutation. Il ne sera donc pas possible par exemple de comparer de manière pertinente deux graphes directement au moyen de leurs matrices d'adjacence.
Comme proposé dans [11], il est tentant de vouloir représenter un graphe par le nombre d'occurences d'une famille de plus petits graphes et cette approche a donné des premiers résultats encourageants. Cependant, la famille de graphes considérée doit être connue à l'avance et posséder de bonnes propriétés (en terme de largeur arborescente) pour que la représentation soit accessible.
L'extension de ces travaux à des familles apprises de petits graphes (dans l'esprit de l'apprentissage de dictionnaire comme illustré en Fig.1) aura pour but de réduire la taille de la répresentation et d'en augmenter son interprétabilité. La formulation de ce problème et sa résolution via des méthodes de programmation mathématique seront investiguées. Cette première piste est à mettre en relation avec des travaux précédents de calcul de graphe moyen [3].

Ces travaux trouve des applications naturelles en chemoinformatique où la structure des molécules est caractéristique de leur comportement et se modélise naturellement par des graphes. Suivant l'intérêt du candidat, des applications en sécurité informatique (où un programme malveillant pourra être modélisé par son graphe d'appel de fonctions) pourront être envisagées.

Collaboration

Ce sujet de thèse fait suite au stage intitulé "Apprentissage de distance d'éditions entre graphes par Réseaux de Neurones", fruit d'une collaboration entre le GREYC, le LITIS et le LAMSADE. Naturellement, les premiers pas de cette thèse ont pour but de prolonger et de dépasser les travaux de l'équipe d'encadrement [3, 6, 8].

Ce sujet de thèse sera effectué en étroite collaboration avec : Benoit Gaüzère (LITIS - INSA de Rouen), Sébastien Bougleux et Luc Brun (GREYC, ENSICAEN).

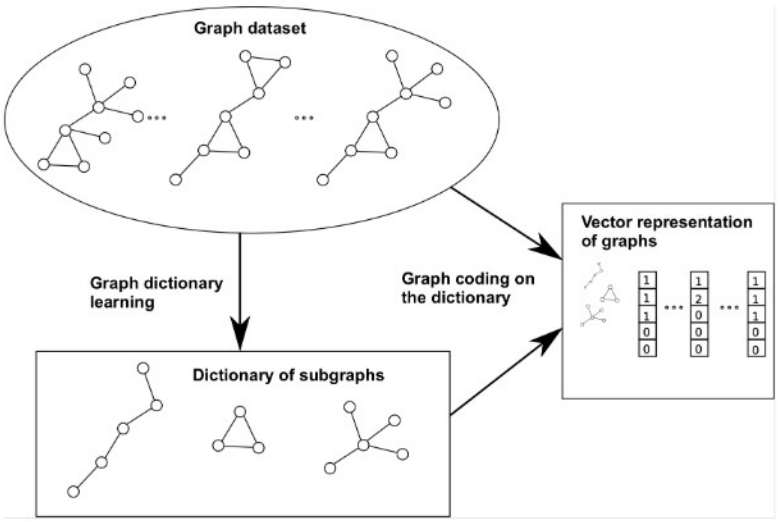

Fig. 1: Apprentissage de dictionnaire de graphes

Profil du candidat

Nous recherchons des candidats et candidates très motivés, titulaires d'un excellent diplôme (master degree, école d'ingénieurs) dans les domaines des mathématiques appliquées ou de l'informatique. Les candidat et candidates doivent pouvoir démontrer d'un solide bagage en apprentissage artificiel et en théorie des graphes et doivent être à l'aise avec le développement informatique.

Processus de sélection

Toute candidature doit comporter un CV détaillé, une lettre de motivation, les relevés de notes d'une ou deux années antérieures et un résumé du mémoire de master.

Références

[1] Aurélien Bellet, Amaury Habrard, and Marc Sebban. A Survey on Metric Learning for Feature Vectors and Structured Data. 2013.

[2] Karsten Borgwardt, Elisabetta Ghisu, Felipe Llinares-López, Leslie O'Bray, and Bastian Rieck. Graph kernels : State-of-the-art and future challenges. *arXiv preprint arXiv :2011.03854*, 2020.

[3] Nicolas Boria, Benjamin Negrevergne, and Florian Yger. Fréchet mean computation in graph space through projected block gradient descent. In *ESANN 2020*, 2020.

[4] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning : going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4) :18–42, 2017.

[5] Hongyun Cai, Vincent W Zheng, and Kevin Chen-Chuan Chang. A comprehensive survey of graph embedding : Problems, techniques, and applications. *IEEE Trans. Knowl. Data Eng.*, 2018.

[6] Lucas Gnecco, Nicolas Boria, Florian Yger, Sébastien Bougleux, and David Blumenthal. Compressing graph collections based on edit arborescences. In *15th Workshop on Compression, Text and Algorithms (WCTA)*. https ://bougleux.users.greyc.fr/articles/05-WCTA_2020_abstract.pdf, 2020.

[7] Martin Grohe. word2vec, node2vec, graph2vec, x2vec : Towards a theory of vector embeddings of structured data. In *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pages 1–16, 2020.

[8] Linlin Jia, Benoit Gaüzère, Florian Yger, and Paul Honeine. A metric learning approach to graph edit costs for regression. *Proceedings of S+SSPR*, 2021.

[9] Nils M Kriege, Fredrik D Johansson, and Christopher Morris. A survey on graph kernels. *Applied Network Science*, 2020.

[10] Andreas Loukas. How hard is to distinguish graphs with graph neural networks ? In *NeurIPS*, 2020.

[11] Hoang Nguyen and Takanori Maehara. Graph homomorphism convolution. In *International Conference on Machine Learning*, pages 7306–7316. PMLR, 2020.

[12] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE Trans. Neural Networks Learn. Syst.*, 2020.