ESFF-GCN : Module d'échantillonnage pour l'entraînement des réseaux de neurones pour graphe

Abderaouf Gacem*, Mohammed Haddad*, Hamida Seba*

*Univ Lyon, UCBL, CNRS, INSA Lyon, LIRIS, UMR5205, Villeurbanne, France
prenom.nom@univ-lyon1.fr

1 Introduction

Le paradigme d'entraînement par lots a été largement adopté par la communauté d'apprentissage profond, notamment pour la vision des ordinateurs et le traitement des langages naturels. Avec l'émergence de l'apprentissage sur les graphes, plusieurs travaux ont tenté d'étendre ce paradigme à l'entraînement des réseaux de neurones sur les graphes (ou Graph Neural Networks GNNs). L'existence de plusieurs travaux est dû au fait que la génération de lots pour les graphes n'est pas aussi simple que la génération de lots pour les autres types de données. Cette difficulté vient d'une part de la dépendance entre les nœuds d'un graphe, et d'autre part par la perte d'information une fois ces nœuds dissociés en lots. Il se trouve aussi que l'entraînement des GNNs par lots apporte un effet d'augmentation de données car le modèle est exposé à différentes captures partielles du même graphe à chaque fois. Cette particularité fait que cette piste de recherche est une piste ouverte malgré l'existence de plusieurs méthodes efficaces pour la génération de lots. En effet, toute contribution de génération de lots apportant une amélioration, s'ajoute à l'arsenal existant autant que collaborateur et non pas comme concurrent. Dans cet esprit, on propose une méthode de construction de lots compatible avec l'entraînement des GNNs. Nos principales contributions sont les suivantes : (1) Nous proposons une méthode d'échantillonnage efficace pour échantillonner des sous-graphes tout en préservant les propriétés du graphe d'origine. Notre méthode d'échantillonnage maintient un niveau de connexité qui permet de ne pas détériorer la qualité d'apprentissage du GNN sur les lots, et elle est suffisamment diversifiée pour permettre au GNN de mieux généraliser lors de l'apprentissage. (2) Nous montrons que notre processus est une généralisation de plusieurs stratégies d'échantillonnage existantes utilisées pour la construction de lots. (3) Nous montrons également que notre processus peut être utilisé avec plusieurs modèles GNN bien connus et que ces modèles s'améliorent en termes de résultats et en termes de capacité à gérer de grands graphes.

2 ESFF-GCN : Échantillonnage de Sous-graphes par le modèle de Feu de Forêt pour l'entraînement de GCN

ESFF-GCN est une méthode d'échantillonnage inspirée du modèle feu de forêt (ou *Fire Forest* FF) utilisé pour la génération de graphes introduite par Leskovec et al. (2005) pour modéliser l'évolution des réseaux du monde réel. Ce modèle préserve les propriétés des graphes

TAB. 1 – *Résultats en pertinence sur les jeux de données*

	Cora	Citeseer	Pubmed	Flickr	Reddit
GraphSAGE	76.11%	66.92%	76.33	46.13%	93.27%
Cluster-GCN	69.17%	63.51%	80.56%	47.44%	95.07%
GCN	80.92%	71.23%	78.60%	46.55%	92.68%
ESFF-GCN$_{faible}$	76.56%	69.93%	74.27%	46.45%	85.66%
ESFF-GCN$_{élevé}$	79.35%	70.66%	78.23%	**47.09%**	83.74%
ESFF-GCN	**81.25%**	**73.98%**	**80.22%**	46.97%	**93.86%**
GAT	**82.94%**	72.56%	78.78%	out-of-meme	out-of-meme
ESFF-GAT$_{faible}$	78.31%	68.24%	75.11%	44.36%	74.45%
ESFF-GAT$_{élevé}$	78.66%	66.17%	73.87%	45.73%	79.89%
ESFF-GAT	80.64%	**73.27%**	**81.65%**	**47.77%**	**85.12%**

du monde réel telles que la densité, les communautés, les distances et la distribution des degrés. Cela a été vérifié par simulation dans (Leskovec et al., 2005). Ce modèle de graphe nécessite un seul paramètre, p, appelé probabilité de combustion ou probabilité de propagation du feu.

Nous nous inspirons de ce modèle pour définir un processus pour échantillonner des sous-graphes en déclenchant et propageant un feu, d'où le nom de processus de propagation en feu de forêt. Le lot est ainsi le sous-graphe G' induit par les nœuds échantillonnés.

Nous avons évalué notre méthode sur la tâche de classifications de nœuds pour des datasets de citations et de réseaux sociaux disponibles publiquement. Nous nous comparons à GraphSAGE qui ne génère pas de lots, et à ClusterGCN qui se base sur le clustering au lieu du sampling pour la génération de lot.Les petites valeurs de p rendent notre méthode équivalente à GraphSaint-RW et les grandes valeurs la rendent équivalente à une construction BFS ou SnowBall. On combine notre processus avec GCN (Kipf et Welling, 2017) et GAT qui est un GCN doté du mécanisme d'attention. Le tableau des résultats montre l'amélioration en pertinence apportée par notre processus et aussi la réduction de mémoire qui permet d'entraîner le GAT. La supériorité de notre processus par rapport aux autres méthodes d'échantillonnage montre que la qualité des lots générés par notre méthode est meilleure sur ces datasets. De plus, pour les datasets où les autres méthodes d'échantillonnage donnent de meilleurs résultats, notre méthode peut s'adapter en agissant sur le paramètre p seulement.

N. B. : Ce travail a été effectué dans le cadre du projet ANR COREGRAPHIE ANR-20-CE23-0002.

Références

Kipf, T. N. et M. Welling (2017). Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Leskovec, J., J. M. Kleinberg, et C. Faloutsos (2005). Graphs over time : densification laws, shrinking diameters and possible explanations. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, Illinois, USA, August 21-24, 2005*.