Sambet Mathys p2207653
Gros Loric p1908872

Rapport BIML : Projet GNN

1 Introduction

La prédiction de liens est une tâche importante dans l'analyse des réseaux. Comprendre quand plusieurs nœuds sont connectés entre eux est un problème pertinent car c'est utilisé dans de nombreux domaines : réseaux sociaux, systèmes de recommandation et beaucoup d'autres.

Nous allons donc utiliser les réseaux de neurones pour pouvoir faire de la prédiction de liens dans des graphes. En particulier les modèles comme les Graph Convolution Networks (GCN) et les Variational Graph Autoencoders (VGAE), qui sont efficaces pour la prédiction de liens.

L'objectif de l'étude est de comparer plusieurs méthodes basées sur les GNN pour la prédiction de liens. On veut cacher un pourcentage de liens dans notre graph et être capable de redécouvrir ces liens avec nos modèles de GNN.

Pour évaluer nos modèles, on utilise les métriques de l'AUC (Area Under the Curve) et de l'Average Precision (AP). Cela permet de mesurer la capacité des modèles à distinguer les liens existants des non-liens, même si certains sont cachés.

On essaie donc d'identifier les meilleures architectures pour la prédiction de liens, mais aussi de comprendre les composants qui influencent la performance des modèles.

2 Modèle de prédiction de liens

Pour la prédiction de liens, nous allons utiliser des approches basées sur les réseaux de neurones graphiques (Graph Neural Networks, GNN). Les GCN et VGAE, ces modèles sont capables d'apprendre des représentations des nœuds du graphe et vont permettre la prédiction de liens.

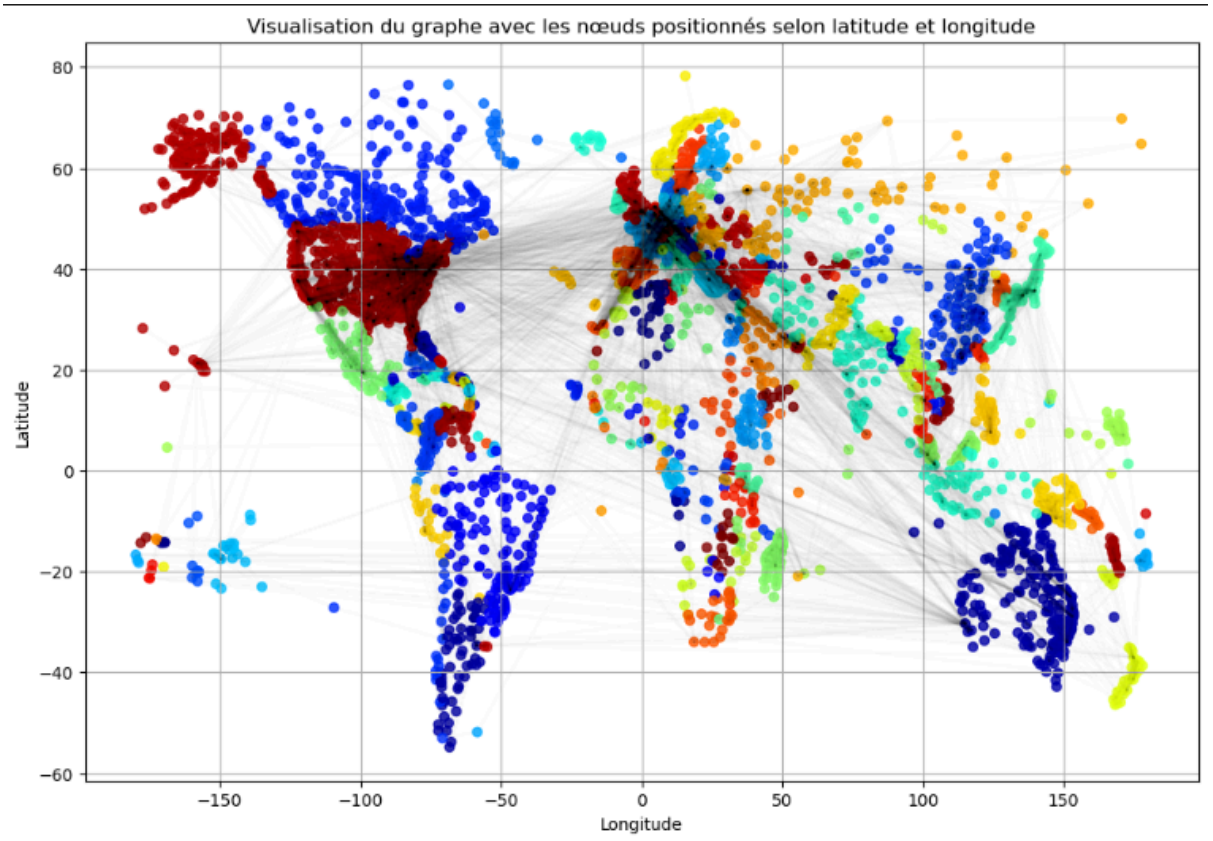
Graph Convolutional Networks (GCN)

1) Variational Graph Autoencoder (VGAE)

Afin d'évaluer les performances des modèles de GNN pour la tâche de prédiction de liens, nous avons utilisé deux métriques classiques : l'AUC et l'AP. Elles sont utilisées dans les tâches de prédiction de lien entre deux nœuds.

3 Protocole expérimental

Pour commencer, le jeu de données utilisé est un graphe composé d'information sur des aéroports dans le monde et chaque nœud est connecté s'il existe un avion qui part du premier nœud pour aller au deuxième.



On modifie ce jeu de test en créant un jeu de validation à partir de ces données pour pouvoir évaluer nos modèles.

La tâche de prédiction de lien consiste à déterminer si un lien existe entre deux nœuds. On a donc fait un masquage des arêtes existantes pour pouvoir appliquer nos modèles dessus et tester leur capacité à redécouvrir les liens cachés.

Dans un premier temps, on cache 10% des liens. On prend aléatoirement 10% des liens dans notre graphe et on les cache. Le modèle doit prédire à partir des 90% restants de liens connus.

Ensuite, on cache 20% des liens. La tâche est plus difficile car moins de données et le modèle doit prédire avec 80% des liens de base.

Une fois les deux tests effectués, on peut faire une comparaison entre les modèles avec l'AUC et l'AP évalué comme base de comparaison.

Pour choisir le meilleur modèle, on teste différents hyperparamètres :

- learning rate entre 1e-5 et 1e-2
- hidden size entre 20 et 200
- nombre d'époques entre 10 et 100

On teste aléatoirement des combinaisons dans ces intervalles et on garde le modèle avec le meilleur résultat.

4 Résultats

Présentation des résultats obtenus pour la prédiction de liens. On analyse la performance de chaque modèle en fonction du pourcentage de liens cachés et des métriques d'AUC et d'AP.

Hyperparamètre des modèles :

GNN 10% de lien cachés : learning rate : 3.538e-05, hidden size : 60, epochs : 3

GNN 20% de lien cachés : learning rate : 1.314e-05, hidden size : 60, epochs : 6

VGAE 10% de lien cachés : learning rate : 0.010, hidden size : 200, epochs : 98

VGAE 20% de lien cachés : learning rate : 0.010, hidden size : 180, epochs : 100

Tableau des résultats

Modèle	10% de liens cachés AUC	10% de liens cachés AP	20% de liens cachés AUC	20% de liens cachés AP
GCN	0.915	0.92	0.903	0.91
VGAE	**0.975**	**0.978**	**0.972**	**0.976**

On voit que VGAE montre de meilleurs résultats sur tous les scénarios qui ont été testés. Le modèle est donc meilleur pour la prédiction de liens manquants et pour pouvoir faire la reconstruction de graphes.

L'impact du pourcentage de liens cachés montre une baisse variable en fonction du modèle dans la prédiction de lien.

5 Ablation study

Pour évaluer l'importance des différentes composantes du modèle, nous avons effectué une étude d'ablation sur le modèle VGAE, qui s'est révélé être le plus performant. Nous avons donc testé les configurations alternatives suivantes pour vérifier l'importance de chaque composant :

- Sans perte KL : En retirant la partie variationnelle de l'auto-encodeur, la performance à diminué pour l'AUC (on passe à 0.96 sans), ce qui montre que la modélisation de l'incertitude permet d'améliorer la prédiction des liens

- Sans couche GCN : En retirant les couches convolutionnelles utilisées pour encoder les nœuds, la performance du modèle a chuté significativement à 0.93 pour l'AUC Et l'AP. Cela montre l'importance de cette couche pour la prédiction.

- Simplification du décodeur : En remplaçant le décodeur probabiliste par un décodeur plus simple (produit scalaire), la performance a augmenté légèrement d'environ 1% pour l'AUC et l'AP. Donc le décodeur probabiliste n'avait pas de raison d'être utilisé car il n'était pas spécialement avantageux.

6 Conclusion

Dans ce projet, nous avons étudié et comparé différentes approches sur les GNN pour la prédiction de liens dans des graphes, en utilisant des modèles GCN et VGAE. Avec les expérimentations effectuées sur le jeu de données, on a pu voir les différents résultats de chaque modèle. Grâce à ces résultats, on a pu remarquer que VGAE se démarque comme étant le plus performant pour la tâche de prédiction de lien après en avoir masqué.

VGAE a montré de bons résultats même lorsque le masquage était de 20% montrant sa robustesse dans différentes situations. GCN a lui montré de moins bons résultats dans l'ensemble. On a aussi observé la diminution de performance lors d'un masquage de lien plus élevé et donc une prédiction de moins bonne qualité.

L'étude d'ablation réalisée sur VGAE a montré l'importance de chaque composant dans le modèle afin de pouvoir prouver leur importance dans la performance globale du modèle.