جامعة دمشق
كلية الهندسة الميكانيكية والكهربائية
قسم هندسة الحواسيب والأتمتة

تطوير خوارزمية لمعالجة الاضطرابات في الشبكات العصبونية الممثلة بالبيان

Developing an Algorithm to Handle the Perturbations of Graph Neural Networks

ضياء حسن هرموش

المشرف العلمي: الأستاذ الدكتور سمير كرمان المشرف المشارك:الدكتورة هيام خدام

الملخص

حققت الشبكات العصبونية الممثلة بالبيان GNN (Graph Neural Networks) نجاحا ملحوظا في العديد من التطبيقات الخاصة بتحليل الرسوم البيانية ونمذجتها.

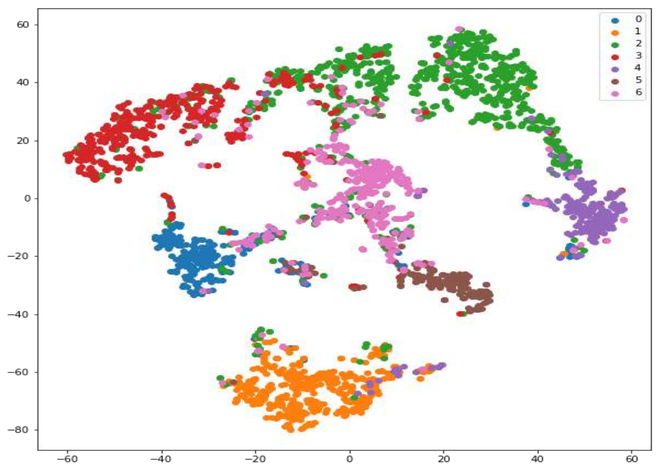
ويعود سر النجاح الكبير الذي حققته GNN في العديد من التطبيقات المتعلقة بالرسوم البيانية الى مخطط تمرير الرسائل الذي تعتمده أثناء التعلم حيث تقوم بتجميع رسائل الجوار لكل عقدة في كل طبقة من طبقاتها أثناء التدريب مما يسمح للنموذج في الطبقة النهائية من معرفة البيان بشكل كامل وفقا للرسائل المجمعة من كل عقدة وجوارها.

تم في هذا البحث دراسة الأثر السلبي للهجمات العدائية على GNN واقتراح نموذج هجوم على البيان المعروف بشبكة الاقتباسات Citation Network لقاعدة البيانات المعروفة بـ CORA-DATSET بعد تحويل هذا البيان الى بيان موزون واقتراح خوارزمية لمعالجة الاضطرابات الناجمة عن هذه الهجمات وذلك وقت اختبار GNN, بعد ذلك تم اجراء دراسة تحليلية لمعاملات النموذج للكشف عن الهجمات العدائية والاستفادة من المعاملات لتطوير الخوارزمية المقترحة ومعالجة الاضطرابات الناجمة عن الهجمات العدائية التي تحدث أثناء تدريب النموذج, كما تم اجراء نوعين من الهجمات (Random attack, metattack) على كل من قاعدتي البيانات (CORA Polblogs) لتظهر النتائج التحسن الملحوظ في أداء GNN وامكانية الخوارزمية من تقليل الآثار السلبية للهجمات.

القسم النظري

- المفاهيم الأساسية حول التعلم الآلي والرسوم البيانية
- الشبكات العصبونية الممثلة بالبيان وآلية تعلمها

القسم العملي

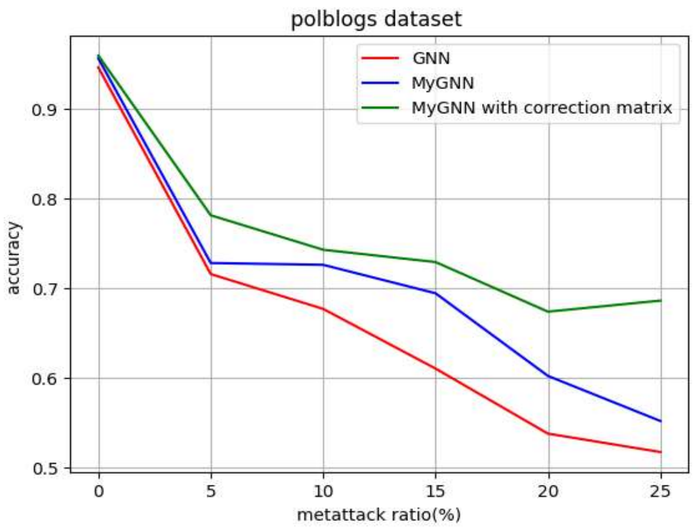


بعد تدريب نموذج GNN على أحد أنواع الرسوم البيانية المصمم للأغراض البحثية قمنا بما يلي:

- ✓ تحويل البيان الى بيان موزون ودراسة الأثر السلبي للهجمات العدائية من خلال تنفيذ هجوم باضافة أو حذف بعض الأضلاع مع تغيير أوزانها أو تغيير بعض الأوزان لأضلاع موجودة مسبقا وفقا لنسب مختلفة
- ✓ اقتراح خوارزمية(خوارزمية التصحيح) لتحسين أداء ومتانة GNN ضد الهجمات التي تحدث وقت الاختبار والتحقق من كفاءة الخوارزمية وفقا لنسب هجوم مختلفة

- ✓ اقتراح أنواع مختلفة من الهجمات التي تحدث خلال تدريب النموذج ودراسة أثرها السلبي على متانة GNN
- ✓ الكشف عن هذه الهجمات بالاستفادة من معاملات الشبكة على نوعين من قاعدتي البيانات
- ✓ اقتراح خوارزمية لمعالجة الاضطرابات التي تنجم عن الهجمات التي تحدث وقت تدريب النموذج واختبار الخوارزمية وفقا لنسب مختلفة
- ✓ تطوير الخوارزمية بالاستفادة من خوارزمية التصحيح واختبارها وفقا لنفس النسب

النتائج والمناقشة



كفاءة النموذج المقترح أفضل مقارنة مع النماذج السابقة من أجل نسب الهجوم الكبيرة

خوارزمية التطوير تمكن من الكشف عن الهجوم وتقليل الآثار السلبية له بغض النظر عن نوعه.

لم نتمكن من مقارنة كفاءة نموذجنا مع النماذج السابقة وفقا للرسم البياني POLPLOGS لأنها تتعامل مع نوع غير موزون بينما تعاملنا مع نوع معدل موزون

لا تحتاج الخوارزمية لمعرفة مسبقة بطبيعة الهجوم

خوارزمية التطوير تمكن من بناء نموذج GNN قوي ومتين ضد الاضطرابات الناجمة عن الهجمات العدائية

المراجع

- [1] H. Dai, H. Li, T. Tian, X. Huang, L. Wang, J. Zhu, and L. Song. Adversarial attack on graph structured data. arXiv preprint arXiv:1806.02371, 2018
- [2] D. Z¨ugner and S. G¨unnemann. Adversarial attacks on graph neural networks via meta learning. arXiv preprint arXiv:1902.08412,2019.
- [3] Hamilton, W. L.Graph representation learning. Morgan & Claypool Publishers, 2020.