Adversarial Attacks on Graph Neural Networks

Presentation of work originally published in the Proc. of the 24th ACM SIGKDD Conference on Knowledge Discovery and Data Mining as well as the International Conference on Learning Representations 2019

Daniel Zügner[1], Amir Akbarnejad[1], Stephan Günnemann[1]

Keywords: deep learning; graph neural networks; adversarial machine learning

Graphs are at the core of many high impact applications ranging from the analysis of social and rating networks (Facebook, Amazon), over gene interaction networks (BioGRID), to interlinked document collections (PubMed, Arxiv). Deep learning models for graphs have advanced the state of the art on many tasks such as node classification or link prediction; they are currently being deployed in production systems, e.g. for content recommendation on social media [Yi18]. Despite their recent success, little is known about their robustness. Yet, in domains where they are likely to be used, e.g. the web, adversaries are common. Can deep learning models for graphs be easily fooled? In [ZAG18, ZG19] we introduce the first studies of adversarial attacks on graph neural networks, aiming to reduce their performance by adding small perturbations to the data. In addition to attacks at test time, we tackle the more challenging class of poisoning/causative attacks, which focus on the training phase of a machine learning model. We generate adversarial perturbations targeting the *node features* and the *graph structure*, thus, taking the dependencies between instances in account. Moreover, we ensure that the perturbations remain *unnoticeable* by preserving important data characteristics. We propose two algorithms: one for *targeted* attacks whose goal is to misclassify a specific target node and one for *global* adversarial attacks aiming to reduce the overall classification performance on test data. The former exploits incremental computations for efficient targeted attacks, and the latter uses meta-gradients to directly tackle the bilevel problem underlying training-time attacks, essentially treating the graph as a hyperparameter to optimize. Our experiments show that small graph perturbations consistently lead to a strong decrease in performance for graph convolutional networks, transfer to unsupervised embeddings, and likewise are successful even when only limited knowledge about the graph is given. Remarkably, the perturbations created by our global attack algorithm can misguide the graph neural networks such that they perform *worse* than a linear classifier that ignores all relational information. Our findings emphasize that further research is needed to improve the robustness of graph neural networks.

[1] Technical University of Munich [zuegnerd,akbarnej,guennemann]@in.tum.de

doi:10.18420/inf2019_29

References

[Yi18] Ying, Rex; He, Ruining; Chen, Kaifeng; Eksombatchai, Pong; Hamilton, William L; Leskovec, Jure: Graph convolutional neural networks for web-scale recommender systems. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, pp. 974–983, 2018.

[ZAG18] Zügner, Daniel; Akbarnejad, Amir; Günnemann, Stephan: Adversarial attacks on neural networks for graph data (*Best Paper Award*). In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 2847–2856, 2018.

[ZG19] Zügner, Daniel; Günnemann, Stephan: Adversarial Attacks on Graph Neural Networks via Meta Learning. In: International Conference on Learning Representations (ICLR). 2019.