OFFRE DE STAGE

Title: Towards randomized algorithms for preserving robustness of graph classification networks against adversarial attacks

Keywords: Neural Networks, Adversarial attacks, Graph classification, Robustness, Randomized algorithms.

Publication Date: 15 October 2018

Beginning Date: February, March or April 2019

Duration: 5 or 6 months

Contact: rafael.pinot@cea.fr and cedric.gouy-pailler@cea.fr

Context

Graphs represent a useful representation for many types of data, widely used in e.g. bioinformatics, network analysis, etc. More broadly, any dataset can be converted into a graph by using a well-chosen similarity matrix construction. In that respect, understanding and analyzing graph structures appears to be a key tool in many real-world applications. It is thus essential to find efficient methods for tasks such as node or graph classification.

In the last decade, deep neural networks have reached an outstanding level of accuracy in numerous areas, such as image classification [KSH12] or object detection [RHGS15]. This models also recently played an important role in node classification and graph classification problems [SGT+09, DDS16, KW16, HYL17]. Despite these advances, deep learning has proven to be susceptible to adversarial attacks. Adversarial attacks [SZS+13] are small perturbations of an input that fools the results of classification for a network. These kind of attacks have also been developed in the context of node classification and graph classification [ZAG18, DLT+18]. Adversarial attacks raise questions of security and safety, and also responsibility in terms of law.

Adversarial examples attacks against machine learning models have become a burning issue due to their efficiency, and the number of sensitive domains they could affect. Accordingly, both attacks and defenses are developed in a tight back-and-forth [GSS14, PMJ+16, PMG+17, DLT+18, SKC18]. Recently, the idea of using randomization in the learning process to ensure robustness against adversarial examples attacks have been successfully used [XWZ+17, MDST18, LCZH17], but no defense for node or graph classification have been successfully presented yet.

Expected work

- Experimental study of state-of-the-art attacks.
- Theoretical approach of defenses.
- Corresponding implementation, and experimentation of defenses.

Required profil

- You are currently in the final year of engineering school or in M2 at the university with specialization in computer science and/or statistics.
- You have a strong background in applied mathematics/computer science (probability, statistics, graph theory).
- You have good programming skills (Python/R, torch/tensorflow, C++).
- Academic research interests you, but also applications to concrete problems.

Conditions
The internship will mostly take place at CEA LIST in Saclay, where you will work with Rafaël Pinot and Cédric Gouy-Pailler, respectively PhD student and engineer-researcher at CEA-LIST. Some meetings will be held with Florian Yger and Jamal Atif, respectively associate professor and professor at Paris-Dauphine. The internship may lead to a PhD thesis.

Access to the CEA is based on a daily bus network service covering the entire Paris region (several buses from Paris in particular). Remuneration (CEA grids) depending on the original curriculum.

References

[DDS16] Hanjun Dai, Bo Dai, and Le Song. Discriminative embeddings of latent variable models for structured data. *CoRR*, abs/1603.05629, 2016.

[DLT+18] Hanjun Dai, Hui Li, Tian Tian, Xin Huang, Lin Wang, Jun Zhu, and Le Song. Adversarial attack on graph structured data. 2018.

[GSS14] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[HYL17] William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. *CoRR*, abs/1706.02216, 2017.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[KW16] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *CoRR*, abs/1609.02907, 2016.

[LCZH17] Xuanqing Liu, Minhao Cheng, Huan Zhang, and Cho-Jui Hsieh. Towards robust neural networks via random self-ensemble. *CoRR*, abs/1712.00673, 2017.

[MDST18] Seyed-Mohsen Moosavi-Dezfooli, Ashish Shrivastava, and Oncel Tuzel. Divide, denoise, and defend against adversarial attacks. *CoRR*, abs/1802.06806, 2018.

[PMG+17] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, ASIA CCS '17, pages 506–519, New York, NY, USA, 2017. ACM.

[PMJ+16] Nicolas Papernot, Patrick D. McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 372–387, 2016.

[RHGS15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[SGT+09] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.

[SKC18] P. Samangouei, M. Kabkab, and R. Chellappa. Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models. *ArXiv e-prints*, May 2018.

[SZS+13] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2013.

[XWZ+17] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan L. Yuille. Mitigating adversarial effects through randomization. *CoRR*, abs/1711.01991, 2017.

[ZAG18] D. Zügner, A. Akbarnejad, and S. Günnemann. Adversarial Attacks on Neural Networks for Graph Data. *ArXiv e-prints*, May 2018.