Adversarial Attack on Graph Classification via Bayesian Optimization Under Hard-Label Black-Box Setting

Group Leader: Yu Zhou
Group Member: Zihao Dong, Guofeng Zhang, Jingchen Tang

- Problem and goal
  - What do you want to solve? Why do you think it is important?
    - Currently, only a limited number of methods exist for attacking a graph neural network model for graph level classification under a black box setting. Among them, we found this work (link), which trains a surrogate model based on Bayesian Linear Regression to generate perturbed graphs, interesting and promising. Note that this model relies on the output logits to compute attack loss. However, in most cases (real world settings), usually we cannot assume we have access to the model output logits when feeding an input graph to the model. Therefore, in order to make the attack more realistic and more generalizable to real world settings, we seek to extend this soft-label black box attack using Bayesian Optimization to a hard-label setting.
  - What results do you expect?
    - In the experiment comparison section of the Bayesian paper we mentioned above, the authors demonstrated that their model has lower test accuracy than all methods they compared to, except in some cases their model is outperformed by white-box Gradient Based methods. Therefore, we expect our model, if formulated correctly, can achieve a similar attack success rate to the soft-label case.
- Data plan
  - What kind of data sets will be used?
    - We use three real-world graph datasets from three different fields to construct our adversarial attacks, i.e., COIL in the computer vision field, IMDB in the social networks field, and NCI1 in the small molecule field.
  - Where and how do you get the data?
    - From four common TU Datasets: IMDB-M, PROTEINS, COLLAB and REDDIT-MULTI-5K
- Solution Plan
  - What is the sketch of the proposed method (related to GNN)?
    - In the original attack, they query the GNN with a perturbed graph and use the loss between original logits prediction and perturbed logits prediction to optimize the feature in Bayesian space to find better perturbation
    - We hope to change this loss optimization to a gradient based method with binary search in black box setting. There might need more thoughts on this, but we are confident that this should work.
  - What will be the major contributions of the proposed method?

        - Firstly, since both Bayesian optimization attacks are query efficient in soft-black box setting, extending it to hard label black box attack can make it closer to the real world and to be more efficient than other hard label black box attacks (which still has plenty of room to work on).
        - Secondly, this might help to challenge some robustness mechanisms of current other GNN models to further improve research in related areas.
- 5. Evaluation Plan
    - How will you evaluate your solution? What evaluation measures are you planning to use?
        - We plan to evaluate our attack method by measuring its attack success rate on representative GNN models using our selected databases.
        - If time permits, we will also consider producing average perturbation(the average number of perturbed edges across the successful adversarial graphs), average queries (the average number of queries used in the whole attack), and average time (i.e., the average time used in the whole attack). An attack has better attack performance if it achieves a larger success rate or/and a smaller average perturbation, average queries and average time.
    - What are the baseline methods?
        - We currently plan to use the following three methods as baseline:(subject to change)
            - Random Attack
            - RL-S2V attack (Dai et al. 2018)
            - A Hard Label Black-box Adversarial Attack Against Graph Neural Networks (Mu et al. 2021)
- 6. Schedule: timeline of your project and workload distribution among team members
    - Work distribution:
        - Experiments Design: Zhou Yu, Jingchen Tang
        - Theoretical Analysis and proofs: Zihao Dong, Guofeng Zhang
        - Reports: work on corresponding parts
        - Presentation: together
    - Timeline:
        - Week5: Study related papers and come up with theoretical and analytical ideas, and start to design experiments.
        - Week6-Week8: finalizing theoretical ideas and proofs, carry out experiments
        - Week9-Week10: Finalizing reports and presentations, get results from experiments.
- 7. References
    - Adversarial Attacks on Graph Classification via Bayesian Optimisation (https://papers.nips.cc/paper/2021/file/38811c5285e34e2e3319ab7d9f2cfa5b-Paper.pdf)
    - Blindfolded Attackers Still Threatening: Strict Black-Box Adversarial Attacks on Graphs (http://yangy.org/works/robust/aaai22_blindfolded_attack.pdf)

- Adversarial Attacks on Graph Classification via Bayesian Optimisation (https://papers.nips.cc/paper/2021/file/38811c5285e34e2e3319ab7d9f2cfa5b-Paper.pdf)