DSC180A Quarter II Written Proposal

Winston Yu, Jianmeng Geng, Barry Xue

Broad Problem Statement

Original: Deep learning models are well known to be vulnerable to adversarial attacks, which are inputs that have been perturbed so minimally such that a human would not misclassify the input but such that a deep learning model would misclassify the input. We propose a new adversarial attack on graph neural networks using variational graph autoencoders and perturbations in the latent space.

Winston's proposed edits: It is now well-known that deep learning models are often vulnerable to adversarial attacks; these attacks make minute changes to an input such that a human would correctly classify the perturbed input, but a deep learning model would not. Current solutions generally protect traditional neural networks and not graph neural networks. With this vulnerability in mind, we propose a new attack on graph neural networks using variational graph autoencoders and perturbations in the latent space.

Detailed Problem Statement

Original: A variational (graph) autoencoder has two components: an encoder, which projects its input into a low-dimensional latent space, and a decoder, which transforms this representation back into the original space that the input inhabits. The detailed problem statement is to optimize a perturbation to some input that a model correctly classifies such that a minimally perturbed representation is decoded into an output that the model incorrectly classifies. In case the proposed adversarial attack does not work (or has already been tried), then we will benchmark its performance compared to other adversarial attacks.

Winston's proposed edits: A variational (graph) autoencoder has two components: an encoder that projects an input into a low-dimensional latent space, and a decoder component, which transforms the latent representation back to the original space of the input. The novel part of our attack is that it perturbs an input in the latent space, as opposed to applying perturbations in the original space. The detailed problem we seek to solve is an optimization problem in which we seek to minimize the latent space distance between the original input and its perturbed version but to maximize the distance between the model's output for the original input and the output for the perturbed input. In case the proposed adversarial attack does not work (or has already been tried), then we will benchmark its performance compared to other adversarial attacks.

Primary Output

Original: The primary output of this project is an innovative adversarial attack on graph neural networks using variational graph autoencoders and perturbations in the latent space. This attack should be able to successfully perturb an input such that a deep learning model misclassified the input, despite minimal perturbations that a human would not misclassify. If the proposed adversarial attack does not work (or has already been tried), then the primary output should be a benchmark of its performance compared to other adversarial attacks.

Winston's proposed edits: The primary output of this project is an innovative adversarial attack on graph neural networks using variational graph autoencoders and perturbations in the latent space. This attack should be able to perturb an input so subtly that a deep learning model would misclassify the perturbed input, but a human would not be able to notice the difference. If the proposed adversarial attack does not work (or has already been tried), then the primary output should be a benchmark of its performance compared to other adversarial attacks

Justification on Success

Datasets we plan to use for method development and benchmarking will be mostly the same as the reference papers.

Datasets used by the reference paper “Graph Robustness Benchmark: Benchmarking the Adversarial Robustness of Graph Machine Learning” are:

1. Cora: 2708 papers, 7 classes, 5429 links, 1433 unique words features.
2. CiteSeer: 3312 papers, 6 classes, 4732 links, 3703 unique words features.
3. Flickr: 105938 images, 2316948 links.
4. Reddit: 232965 posts, 114618780 links, 300 unique words features.
5. Aminer: 29814 papers, 632752 links, each node has 5 features.
6. PubMed: 19717 papers, 44338 links, 500 unique words features.

We will start with the datasets that are readily available to us, which are Cora, PubMed and CiteSeer. The quality of all the dataset is no issue because all datasets’ usage can be traced to existing papers that contain replicable results. Overall, there is no shortage of high quality dataset on which we can evaluate our attack upon.

DSC180A Quarter II 6 Weeks Schedule

Winston Yu, Jianmeng Geng, Barry Xue

Schedule

Week	Task
Winter Break	Review literature, get familiarized with the topic
Week 1	Collect Data
Week 2	Formulating/Implementing new attack method
Week 3	
Week 4	Evaluate and collect results
Week 5	Compare with other approaches
Week 6	Report & Beyond

Responsibility

Week	Barry	Jianming	Winston
Week 1	Collect 2 Dataset; implement data loading; report data statistics	Collect 2 Dataset; implement data loading; report data statistics	Collect 2 Dataset; implement data loading; report data statistics
Week 2	Implement Models/Attacks from paper	Implement Models/Attacks from paper	Formulate and Implement New Attack
Week 3			
Week 4	Run assigned models; and compare results	Run assigned models; and compare results	Run assigned models; and compare results
Week 5			
Week 6	Work on corresponding sections of the report; assigned during Week 5.	Work on corresponding sections of the report; assigned during Week 5.	Work on corresponding sections of the report; assigned during Week 5.