Research Project for CS471

Finding correlation between neighbor overlap and over-smoothing

20190272 박준현
20190605 정인호
20220554 이태윤

Problem Statement

Motivation

- Over-smoothing, which highly depends on the number of layers, degrades the performance of GCN.

- Experimenting with all possible configurations is time-consuming and costly.

- Can we predict the optimal number of layers to avoid over-smoothing and achieve the best performance?

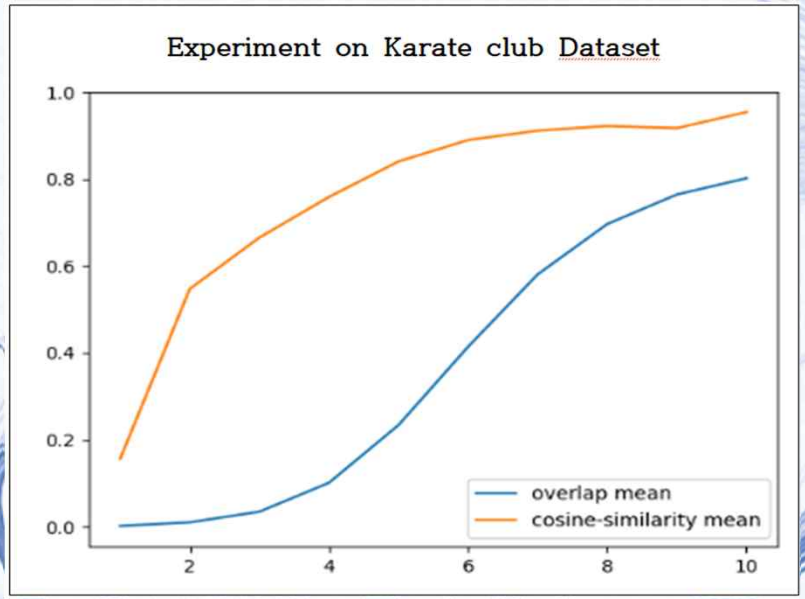
Objective

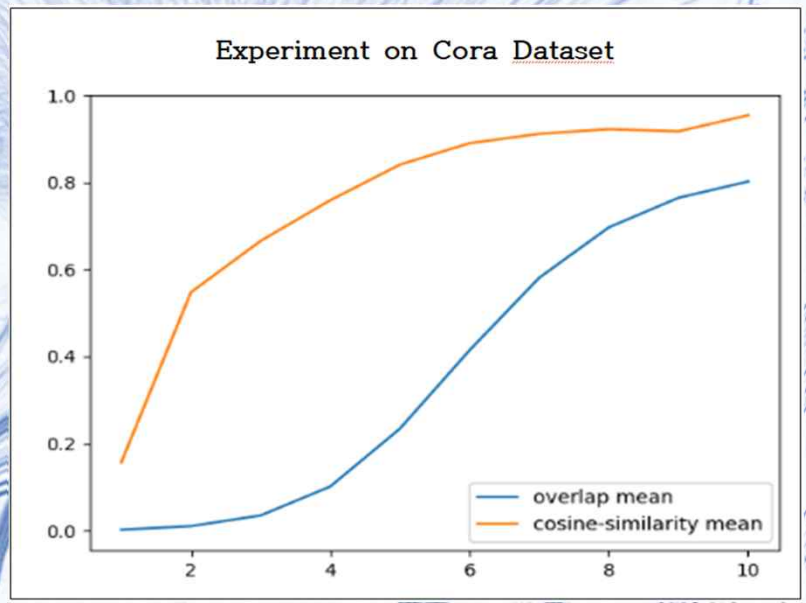
- Predict the number of layers in GCN to avoid over-smoothing and lead its best performance.

High level idea

- Let GCN have k layers. Every node v is affected by its k-hop neighborhoods, $N_k(v)$.
- Define Overlap = $\frac{|N_k(A) \cap N_k(B)|}{|N_k(A) \cup N_k(B)|}$ for any node A, B
- Hypothesis : Increase of $|N_k(A) \cap N_k(B)|$ leads over-smoothing, while increase of $|N_k(A) \cup N_k(B)|$ means more uncommon neighbors. Thus, increase of overlap yields increase of over-smoothing.
- Over-smoothing means the tendency of every node feature representation becomes similar, so we measure feature vector cosine-similarity.
- Want to Find : the correlation between node neighbor overlap and over-smoothing of feature vectors.
- Experiment Design : With various graph datasets such as Karate and Cora, measure the mean of overlap and the mean of cosine similarity with differing layer number k.

Experimental Results





- Karate Club dataset is dense(density = 0.14), while Cora is much sparse(density=0.001)
- We found the mean of overlap and cosine-similarity follows similar tendency in dense graph.
- While in sparse graph, the mean of neighbor overlap grows much slower than cosine-similarity.
- Further research: conduct experiments on various graphs with different edge density.