Machine Learning Seminar

Self Supervised Learning for Graph Classification

Selim Yagci

Supervisor: Tom Wollschläger

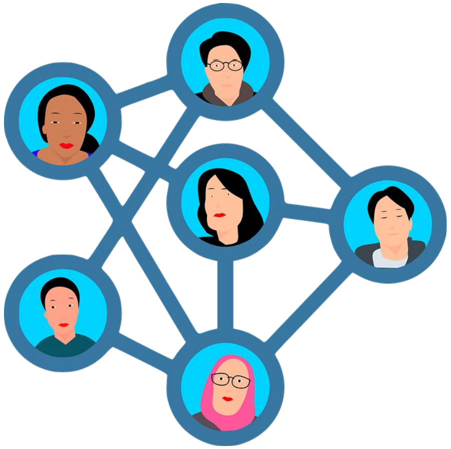
Department of Informatics, Technical University of Munich

18.07.2023

Contents

Why graph-structured data?

- Social networks
- Molecular sciences
- Transportation
- E-commerce



What is wrong with supervised setting?

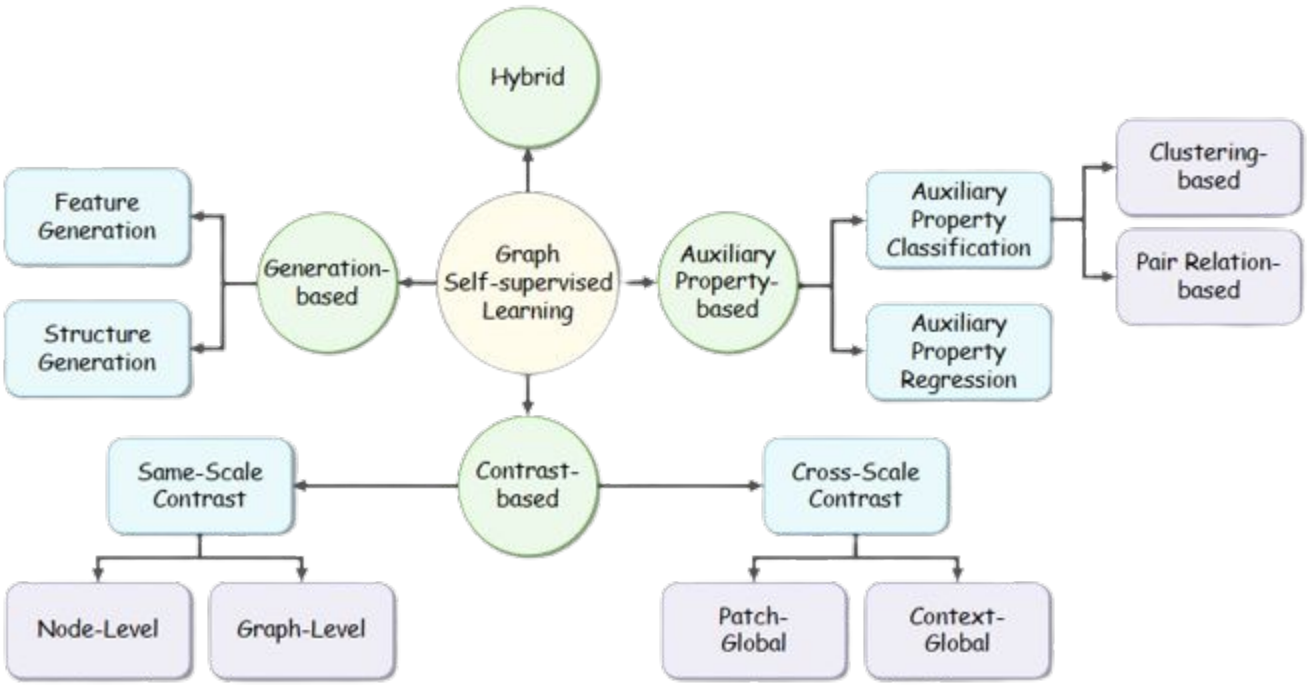
Reliance on labeled data

Poor generalization and less robustness

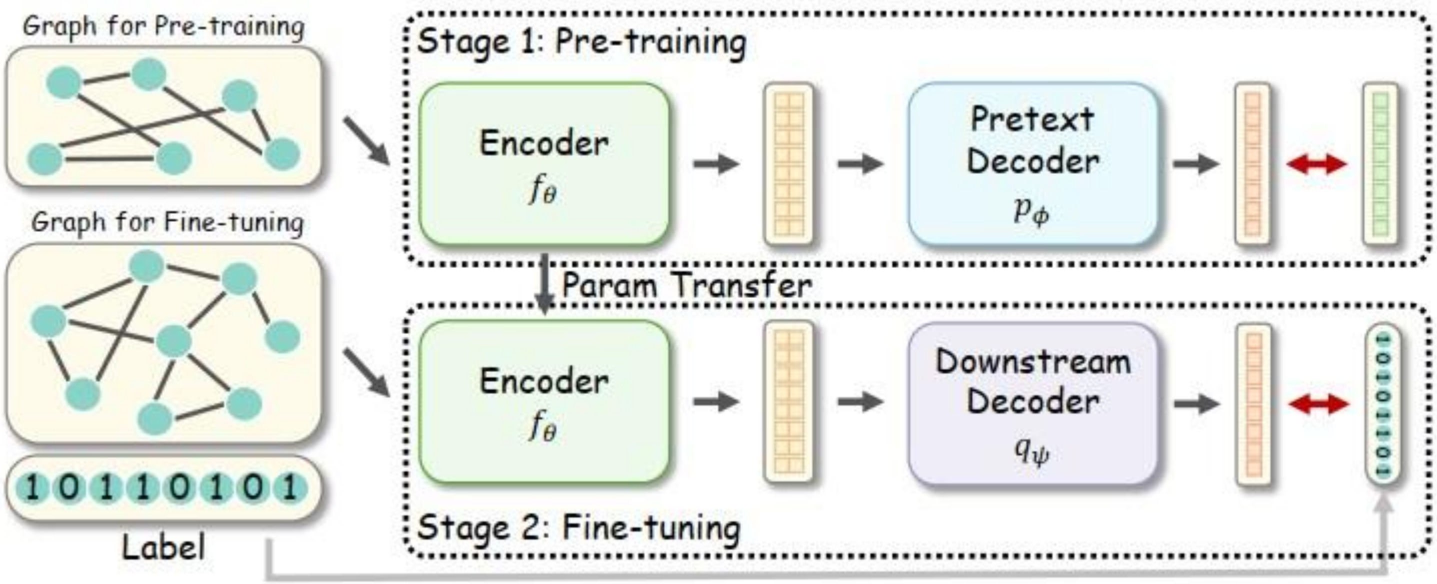
Graph Self Supervised Learning

Pretext vs Downstream Task?

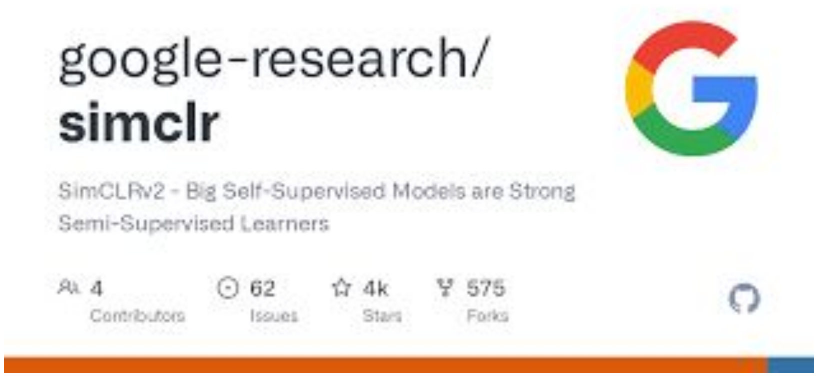
Designing the pretext task



From pretext task to downstream task



Is SSL successful for all?



What about Graphs?

Out-of-distribution Generalization

- In-distribution assumption
- Distribution shift on node features, graph size and other structural properties

How SSL can address these?

Overview of Existing Works

Test Time Adaptation Strategy

To adapt models based on test samples in the presence of distributional shifts

Test Time Training	Fully Test Time Adaptation
Optimizing self supervised auxiliary task jointly with main task on source data	Batch normalization statistics
Fine-tuning the model on test data	Prediction entropy minimization
	Classifier adjustment

Limitations of prior works on graphs

From visual domain to graphs

- Complicated graph augmentations

Data agnostic models

- Entropy Minimization, trusting false pseudo-labels leads to confidence bias

Self supervised auxiliary task (contrastive learning)

- GNN encoders may learn label irrelevant redundant information through

A tailored method for graphs?

Graph Adversarial Pseudo Group Contrast (Chen et al., 2022)

Group Pseudo-Positive Samples (GPPS)

- Contrastive Learning variant over Entropy Minimization
- Mini batch of positive samples over single positive sample
- Maximizing similarity between anchor graph and positive samples
- More informative pseudo-labels
- Less confidence bias
- GPPS is suitable for graph CL since it is always label-sensitive

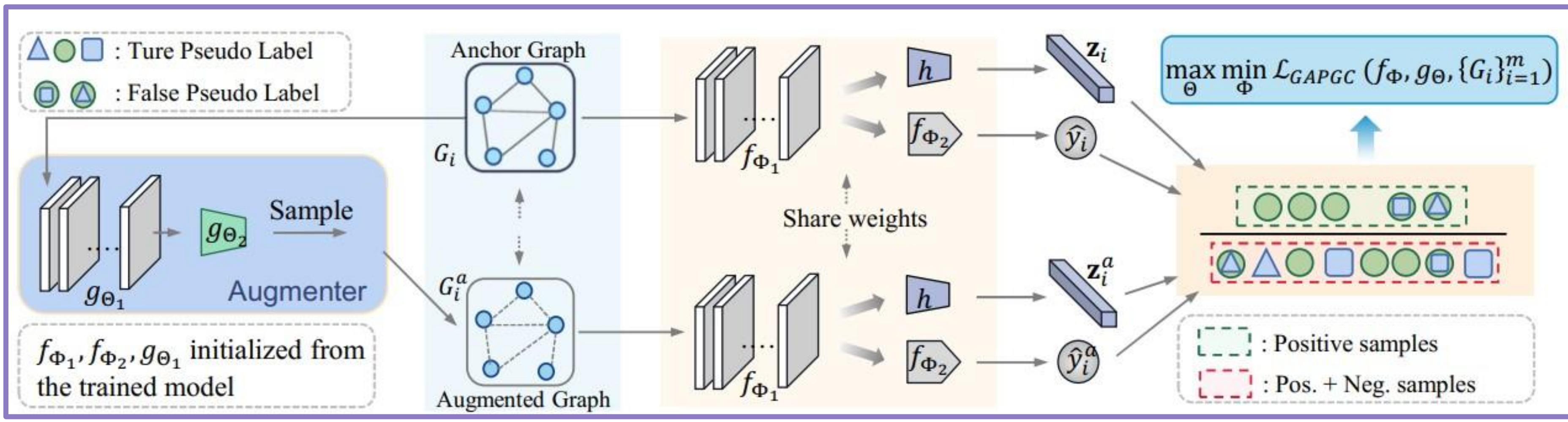
Adversarial Learnable Augmenter (ALA)

- CL can capture redundant info
- Maximizing contrastive loss
- Learnable edge-dropping augmentation
- Parametrized initially from offline trained model
- Less information encoded

Together with GPPS and ALA, GAPGC follows a min-max principle

- Encoder **f** to pull anchor and pseudo-pos samples together and pushing anchor away from neg samples

- Augmenter **g** to max contrastive loss



Performance and problems of the proposed method

- Contrastive learning as self supervised task during testing improves test performance
- ALA contributes more than GPPS but when combined performs the best on average
- Experiments on molecular scaffold OOD dataset demonstrated state-of-the-art performance on GNNs

- Access to test distribution in real datasets?
- Can the parametrized augmenter generalize to other learnable graph generators?
- GCL projector and MLP in augmenter initialization?

From local structures to size generalization (Yehudai et al., 2021)

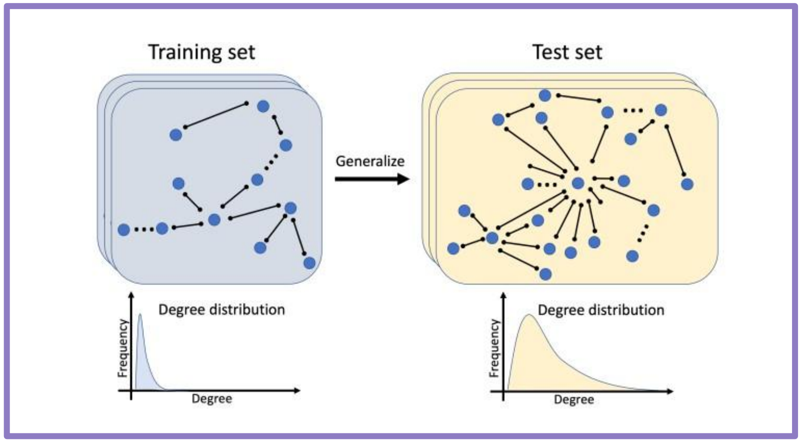
Different size graphs in the same domain

- Few atom compounds to thousands nodes proteins, social networks with dozens of nodes to billions

Hard to collect labels for large graphs

The type of graphs that the distribution of its local structures depends on size

- Preferential attachment in social networks



Graph distribution and local structures

G(n,p) graphs (or Erdos-Renyi)

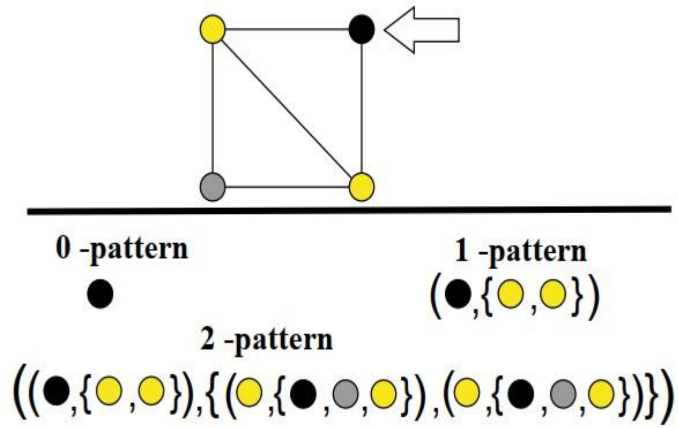
- n nodes and p probability edges
- Increasing n changes local structure by changing node degrees

Preferential attachment

- n nodes and each new node connects to m other nodes with a probability proportional to their degree
- High degree nodes have high chance to connect to new node
- Increasing size also increases maximum degree in the graph and changes local structure

D-pattern Characterization and Expressivity of GNNs

- Inspired by Weisfeiler-Lehman test
- D-pattern of a node is an encoding of d-1 patterns of itself and its neighbors
- The 1-pattern of each node is its degree
- Controlling values of d-layer GNNs on set of d-patterns, which can determine GNNs output
- As d-pattern discrepancy grows, the generalization of GNNs degrades more



How can we exploit this relation?

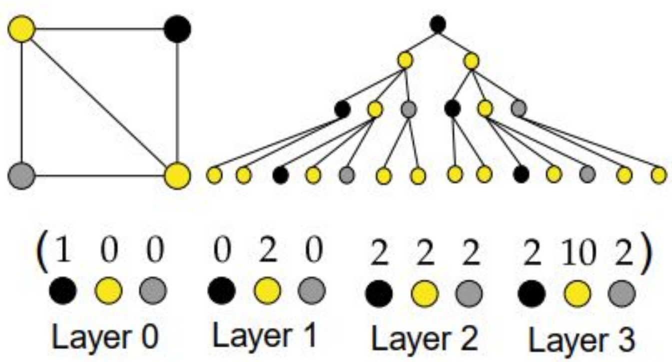
A self supervised domain adaptation method

Setting: Smaller size graphs in source domain, larger ones in target

Pretext task: Learning useful d-pattern representation ~ predicting the node labels holding important information about the node's d-pattern

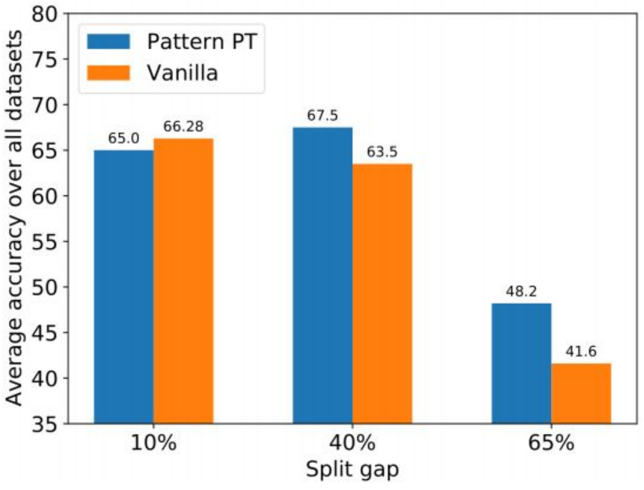
- Constructing the pattern-tree
- Calculate a descriptor of the tree by concatenation of histograms of different node features in each layer
- Train network in a node regression setup to predict this descriptor on both source and target domains

Result: Representations of source and target domains get aligned



Performance and problems of the proposed method

- Performs well on mostly social domain, i.e. IMDB, Twitch, Deezer etc.



- In biology domain performs worse than SOTA and other domains are unknown
- Categorical node feature and bidirectional edges with no feature assumption
- Accessing test distribution

Conclusion and Future Research

- Image and text based vs. Graphs
- Heavy label reliance, poor generalization, and consequently weak robustness
- Extract knowledge through specific pretext tasks without having manual labels, resulting in better generalization

- Reliance on data and model specific assumptions or requirement of accessing test distribution
- Explainability

Future Direction

Automatically finding best suitable pretext task and objective functions for different domains, while preserving the robustness to size shifts (Jin et al., 2022)

Thank you for listening!

Q&A

References

[1] Chen, G., Zhang, J., Xiao, X., and Li, Y. Graphtta: Test time adaptation on graph neural networks, 2022.

[2] Jin, W., Liu, X., Zhao, X., Ma, Y., Shah, N., and Tang, J. Automated self-supervised learning for graphs, 2022.

[3] Yehudai, G., Fetaya, E., Meirom, E., Chechik, G., and Maron, H. From local structures to size generalization in graph neural networks, 2021