Impact of Graph Representation on Machine Learning with Graph Neural Networks

Why

Machine learning on knowledge graphs is becoming a popular area of research which studies how statistical pattern recognition approaches, such as (deep) neural networks, can be applied to complex graphs with rich semantics. A recent development in this field involves the use of graph neural networks, which can learn directly over the entire graph in an end-to-end fashion. To do so, it is commonplace to represent a knowledge graph by a collection of adjacency matrices, one for every unique relation, in which each point (i, j) conveys whether two nodes i and j are connected. Here, no distinction is typically made between nodes that represent *entities* (things, people, etc.) and those that represent the entities' attributes (i.e. literals). Instead, literals with the same value are collapsed into a single node. An example of this are people with the same first name, who are now linked together via the single node that represents that name (see Fig. 1). This unintentionally adds new information to the graph, which impacts the performance of the machine learning process. However, the more faithful representation, which gives each literal its own node, results in much larger adjacency matrices, which also impacts performance, albeit differently.

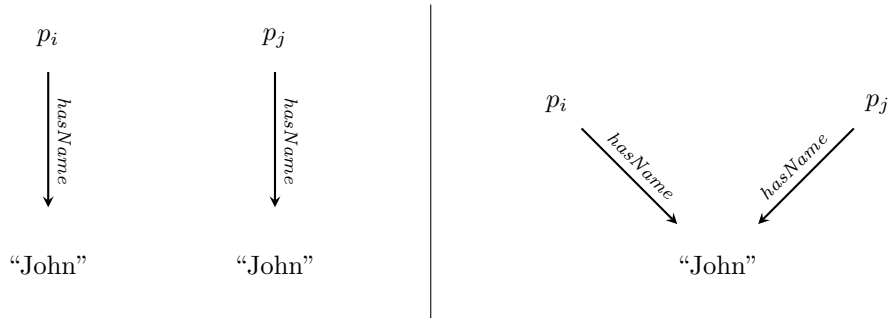

Figure 1: Original graph (left) and with collapsed literals (right)

What

The goal of this project is to evaluate the impact of the choice of graph representation on the performance of a graph neural network, specifically the representation of the graph's literals.

How

This project will follow the follow steps:

- A brief literature review on machine learning on graphs
- Develop a machine learning pipeline to run various graph neural networks
- Run experiments with different graph representations as input
- Evaluate the impact and it compare to related work

Who

This project aims at students who are interested in deep learning on complex structures. A strong familiarity with programming and machine learning is required.

Supervisors

Xander Wilcke & TBA