Lehrstuhl für Angewandte Informatik X

UNIVERSITÄT BAYREUTH

Prof. Dr. Ruben Mayer

Titel

Analyse des Einflusses von Sampling-Strategien im Mini-Batch-Training von Graph Neural Networks auf Modellgüte und Effizienz (B.Sc./M.Sc.)

Hintergrund

GNNs haben sich als leistungsfähige Methode zur Verarbeitung relationaler Daten etabliert. Bei großen Graphen ist Full-Graph Training jedoch oft nicht skalierbar, weshalb Mini-Batch-Verfahren mit Sampling-Strategien eingesetzt werden.
Diese Strategien bestimmen, welche Teile des Graphen während des Trainings berücksichtigt werden, und beeinflussen somit sowohl die Effizienz als auch die Qualität der gelernten Repräsentationen. Die konkreten Auswirkungen unterschiedlicher Sampling-Ansätze auf Modellgüte, Konvergenzverhalten und Trainingsstabilität sind jedoch bislang nicht vollständig geklärt und sollen im Rahmen dieser Arbeit systematisch untersucht werden.

Aufgabenstellung

Ziel der Arbeit ist die systematische Untersuchung, wie unterschiedliche Sampling-Strategien im Mini-Batch Training von GNNs die Modellperformance, Konvergenz und Effizienz beeinflussen, sowie ggf. deren Skalierbarkeit in verteilten Trainingsumgebungen.

Theorieteil:

- Grundlagen zu GNNs und Skalierung großer Graphen beim Training
- Überblick über Sampling-Strategien (Neighbor Sampling (z. B. GraphSAGE), Layer-wise Sampling, Subgraph-basierte Methoden (z. B. Cluster-GCN, GraphSAINT) usw.)
- Analyse theoretischer Eigenschaften (Bias, Varianz, Informationsverlust)

Implementierung:

- Implementierung ausgewählter Sampling-Strategien (PyTorch Geometric oder DGL)
- Durchführung von Experimenten auf Benchmark-Datensätzen
- Vergleich hinsichtlich: Modellgüte (Accuracy, F1, etc.), Trainingszeit und Speicherverbrauch, Stabilität und Konvergenzverhalten
- Systematische Variation von Sampling-Parametern (z. B. Nachbarn, Tiefe)

Option:

Theorie:

- Implikationen von Sampling-Strategien für verteiltes Training: Datenlokalität, Partitionierung von Graphen, Kommunikationskosten

Implementierung:

- Erweiterung der Experimente auf verteiltes Training
- Analyse des Einflusses von Sampling-Strategien auf Kommunikationsaufwand, Trainingszeit im verteilten Setup und Skalierbarkeit

Erforderliche Kenntnisse und Fähigkeiten

Python, PyTorch, DGL, Grundlagen von GNNs und Deep Learning

Betreuerin

Barbara Hoffmann
Fragen jederzeit gerne via Teams oder E-Mail (barbara.hoffmann@uni-bayreuth.de)