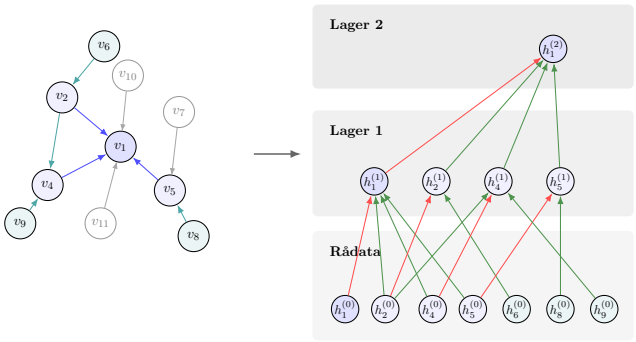
INSTITUTIONEN FÖR DATAVETENSKAP | LUNDS TEKNISKA HÖGSKOLA | PRESENTERAD 2026-06-10
EXAMENSARBETE Cost-Efficient Graph Database-backed Graph Neural Network Pipelines
STUDENTER Victor Pekkari, Harry Wik
HANDLEDARE Xuan-Son Vu (LTH), Brian Shi (Neo4j), Alfred Clemedtson (Neo4j)
EXAMINATOR Jacek Malec (LTH)

Kostnadseffektiv maskininlärning på grafer med grafdatabas-integration

POPULÄRVETENSKAPLIG SAMMANFATTNING **Victor Pekkari, Harry Wik**

Grafdatabaser utgör en växande och effektiv lagringslösning för sammanlänkade dataobjekt. Samtidigt har en anpassning av neurala nätverk som agerar på grafdata blivit en av de mest ledande metoderna för representationslärning av grafer. Vårt examensarbete utforskar hur dessa nätverk effektivt kan integreras med grafdatabaser.

Grafdatabaser är optimerade för att lagra och modellera data som nätverk av noder och relationer. Information kan sparas direkt på både noder och relationer varefter datamängden kan sökas strukturerat. Parallellt har neurala nätverk anpassats för grafstrukturer. Vi studerar en populär variant av dessa: *message-passing neural networks*. Ett sådant system fungerar genom att varje nod skickar en representation av sin information till sina länkade grannar. Varje nod aggregerar sedan sin egen och de inkommande representationerna för att uppdatera sin information. På så vis införlivar noden uppgifter från sitt grannskap. Processen kan upprepas så att information från alltmer avlägsna noder används för att lösa nätverkets optimeringsproblem.



Figur 1: Sampling av nod v_1:s grannskap, följt av beräkning av representationen for nod v_1.

Ett problem är dock att informationsmängden för att processa en enskild nod snabbt kan explodera till att omfatta nästan hela grafen. Därför undersökte vi samplingstekniker som approximerar nodens omgivning och sätter ett övre tak för mängden data som behöver bearbetas.
I vårt arbete utfördes nod-klassificering, vilket innebär att avgöra till vilken kategori en given nod hör. Som tillämpning studerade vi bl.a. kategorisering av vetenskapliga artiklar i citeringsnätverk. Modellen tränades på nyckelord i artiklar inom fält som datavetenskap och ekonomi, men drog även slutsatser baserat på ordval i citerande artiklar. Efter träning kan modellen extrapolera och kategorisera stora mängder data som inte ingått i träningsdatan och manuellt knutits till en kategori.
Det nyskapande i vårt arbete var undersökningen av hur dessa neurala nätverk kan integreras med grafdatabaser. Genom att implementera och jämföra olika integrationstekniker utvärderade vi deras för- och nackdelar. Vi visade att våra metoder för sampling i en grafdatabas var i sin funktion statistiskt oskiljbara från standardimplementationer i maskininlärningsbiblioteket PyTorch Geometric. Att skifta från en RAM-baserad datakälla till en RAM/disk-hybrid via en grafdatabas gör att mindre och billigare maskiner kan användas för träning. Nackdelen är att flödet riskerar att bli långsammare. Vi utvecklade därför ett databasplugin med skräddarsydda funktioner för sampling och kompakt datarepresentation, vilket reducerade träningstiden flerfaldigt. Genom att flytta beräkningar till databasen minimerades dataöverföring till maskininlärningsprocessen.
Slutligen undersökte vi om inferens för en färdigtränad modell med fördel kunde köras helt i databasen. Samtliga implementeringar av inferens i databasen gav minst lika träffsäkra resultat som inferens med PyTorch Geometric. Vidare fann vi att en av dessa metoder var snabbare än att skicka data för beräkning i PyTorch Geometric.