INSTITUTIONEN FÖR DATAVETENSKAP | LUNDS TEKNISKA HÖGSKOLA | PRESENTERAD 2026-06-11

EXAMENSARBETE Iterative Tensorwise Quantization of Neural Networks Using MLIR
STUDENTER Albin Nyström, Aron Somi
HANDLEDARE Noric Couderc (LTH), Felix Malmsjö (Inceptron)
EXAMINATOR Flavius Gruian (LTH)

Iterativ kvantisering av AI-modeller

POPULÄRVETENSKAPLIG SAMMANFATTNING **Albin Nyström, Aron Somi**

Storleken på AI-modeller växer i rasande fart, vilket skapar ett ökat behov av resursoptimering. En metod för storleksminskning av AI-modeller är så kallad kvantisering. De flesta metoderna utför endast kvantisering i ett steg, i vårt arbete utforskar vi möjligheten att göra detta till en iterativ process.

I takt med att dagens AI-modeller blir allt större krävs allt mer av hårdvaran som kör dem, vilket gör tillhandahållandet av AI till ett dyrt och på sikt ohållbart åtagande. Ett sätt att göra AI mer resurseffektivt är att komprimera modellerna.
Vid komprimering använder man sig av datorers sätt att representera siffror. En dator representerar en siffra med ett antal bitar som är antingen `1` eller `0` och den har i sitt minne plats för ett begränsat antal bitar. Ett vanligt sätt frigöra minne är att gå från ett talsystem med hög precision som använder många bitar till ett med färre bitar och lägre precision. Detta kallas kvantisering.
Förenklat är en AI-modell en process som lagrar kunskap i en stor mängd siffror som kallas vikter. Vikterna utgör majoriteten av minnet en modell tar upp, därför är kvantisering av en modells vikter ett effektivt sätt att komprimera modellen. Kvantiseringen har dock ett pris, minskningen i storlek innebär också en minskning i precision vilket i sin tur påverkar prestandan. Alla vikter i en modell har inte lika stor betydelse för dess prestanda. Moderna kvantiseringsmetoder utnyttjar detta och använder olika verktyg för att bedöma vilka delar av en modell som är mer eller mindre lämpliga för kvantisering. Metoderna skapar sedan en s.k. kvantiseringskonfiguration som beskriver hur modellens olika delar ska kvantiseras.
Vårt examensarbete utforskar en metod som steg för steg söker efter den optimala kvantiseringskonfigurationen av en AI-modell. De etablerade kvantiseringsmetoderna utför avancerade analyser av modellen innan kvantisering, men är inte medvetna om resultaten av sin kvantisering. Vi använder oss av en optimeringsmetod som testar en konfiguration, utvärderar den och sedan testar en ny konfiguration baserat på resultaten av den föregående. På sikt "lär" sig metoden vad som gör en konfiguration bra respektive dålig och strävar på så sätt efter en optimal lösning.
Resultatet visar på att vår optimeringsmetod fungerar och att den successivt hittar allt bättre konfigurationer. Vår metod hittar däremot inte konfigurationer som är bättre än de som moderna kvantiseringsmetoder levererar. Vi bedömer att rymden av möjliga konfigurationer som optimeringsmetoden utforskar är för omfattande för att rymmas inom ramen för vårt examensarbete. Vi lämnar därför åt framtida arbeten att vidare experimentera med och optimera denna metod.