EXAMENSARBETE Quantization Techniques for Memory-bound Transformers on Ethos-U Hardware Accelerators
STUDENTER Märta Holmquist, Emma Kujala
HANDLEDARE Flavius Gruian (LTH)
EXAMINATOR Sven Robertz (LTH)

Kvantisering + rotationer = sant?

POPULÄRVETENSKAPLIG SAMMANFATTNING **Märta Holmquist, Emma Kujala**

Det blir alltmer vanligt att stora och beräkningstunga maskininlärningsmodeller ska köras lokalt på väldigt små enheter. *Kvantisering* används för göra dessa modeller mindre, snabbare och mer energieffektiva utan att tappa för mycket precision. Detta arbete undersöker ett sätt att förbättra kvantisering genom att introducera rotationer.

Många kraftfulla AI-modeller körs i dag i molnet, men i en del fall kan exempelvis krav på svarstid, integritet eller tillgång till nätverk göra det fördelaktigt att köra modellen direkt på en enhet. Detta kallas edge-inferens och kan utnyttjas i bland annat fordon, mobiler, kameror, sensorer och annan inbyggd elektronik.

Utmaningen med detta är att små edge-enheter har begränsat minne, låg energibudget och mindre beräkningskraft. Därför behöver modellerna optimeras innan de kan köras på sådan hårdvara. En vanlig metod för detta är kvantisering, där modellens tal approximeras till färre bitar i enklare, diskreta format. På så sätt minskas både minnesanvändning och energiförbrukning, men med potentiella uppoffringar i modellens noggrannhet.

Kvantisering fungerar sämre när modellens aktiveringar innehåller enstaka mycket stora värden, så kallade outliers. Dessa extremvärden kan tvinga kvantiseringen att täcka ett större intervall, vilket gör att de vanligaste värdena får färre tal i intervallet att representeras av och därför avrundas mer. För att minska detta problem kan man använda rotationer. En rotation ändrar inte informationen i modellen, utan uttrycker den i ett nytt koordinatsystem - på så sätt kan extremvärden spridas över flera dimensioner i stället för att dominera en enskild kanal. Det kan göra aktiveringarna jämnare och enklare att kvantisera utan att modellens noggrannhet försämras lika mycket. Olika rotationer är olika bra på att åstadkomma detta, och i detta arbete har vi undersökt slumpmässiga rotationer samt försökt hitta optimala rotationer med hjälp av maskininlärningsalgoritmer.

Vi har även undersökt effekterna av att endast kvantisera delar av modellen, då vissa komponenter är mer känsliga för minskad precision, samt hur detta kan kombineras med rotationer för att balansera minnesanvändning, hastighet och precision på Arm Ethos-U NPU:er.

Resultaten visar att rotationer och partiell kvantisering kan förbättra modellens noggrannhet till nivåer mycket nära den icke-kvantiserade modellen, men till vissa kostnader. Rotationer medförde fler beräkningar och ökad minnesanvändning, medan delvis kvantisering introducerade extra kommunikation mellan externt minne och beräkningsenheter. Rotationer funna via maskininlärning var resurskrävande att ta fram, men kan potentiellt ersättas med slumpmässiga rotationer då dessa i genomsnitt hade en tydligt positiv effekt. Sammantaget visar detta att teknikerna kan vara effektiva men mest användbara i situationer där förbättrad precision motiverar en ökad kostnad i beräkning, minne och svarstid.