INSTITUTIONEN FÖR DATAVETENSKAP | LUNDS TEKNISKA HÖGSKOLA | PRESENTERAD 2020-06-11

EXAMENSARBETE Quantization Profiler for Artificial Neural Networks
STUDENTER Martin Lindström, Jakob Hök
HANDLEDARE Jörn Janneck (LTH), Axel Berg (ARM Sweden AB), Kevin Wohnrade (ARM Sweden AB)
EXAMINATOR Flavius Gruian (LTH)

Optimeringsverktyg som effektiviserar AI på inbyggda system

POPULÄRVETENSKAPLIG SAMMANFATTNING **Martin Lindström, Jakob Hök**

Optimering av slutledning för artificiella neurala nätverk är ett måste vid exekvering på inbyggda system på grund av begränsad datorkraft. Examensarbetet har utvecklat ett verktyg som ska hjälpa utvecklaren att analysera avvägningen mellan prestandaoptimeringar och slutledningsförmåga.

Maskinlärning och AI etablerar sig allt mer i samhället, där de används till att ge videorekommendationer, vänförslag på sociala medier, självkörande bilar, röstassistenter och mycket mera. “Artificiella neurala nätverk“ (ANN) är ett vanligt förekommande uttryck som är en mjukvaruteknik som använder sig av nätverk med moduler. Dessa nätverk tränas för att kunna känna igen och identifiera domänspecifika mönster. Efter träning används de i flera användingsområden, däribland bildklassificering. Vid bildklassificering skickas en bild in till nätverket som sedan gör en kvalificerad gissning på vad som visas på bilden; om exempelvis nätverket har tränats för att kunna särskilja hundar ifrån katter så skulle en modul till exempel kunna ha lärt sig hur en nos ser ut och en annan modul hur en svans ser ut hos de två olika djuren. Efter denna typ av igenkänning tas ett beslut av nätverket om det faktiskt är en hund eller en katt. Fundera på hur du själv skiljer på en hund och en katt!

Att skicka in en bild till ett ANN för att få ett utsägande kallas *slutledning*. Eftersom forskningen kring ANN har ökat explosionsartat det senaste decenniet, mycket tack vare tillgången på stor mängd träningsdata och kraftfullare datorer, har efterfrågan på tekniken ökats. Dessutom finns det ett behov av att kunna behärska tekniken på maskiner med begränsad datorkraft, så som mobiltelefoner. Slutledning kräver mycket datorkraft och därav för att köra det på begränsad hårdvara, behöver utvecklare optimera nätverken till att bli snabbare och effektivare, och samtidigt bibehålla tillräckligt bra slutledningsförmåga.

Ett sätt att optimera ett ANN för en mobilprocessor är att utföra beräkningarna enbart med heltal istället för decimaltal, eftersom decimaltal är mer mer krävande att hantera för en dator. Vad som är intressant för en utvecklare är att enkelt kunna testa och analysera hur olika genvägar och optimeringar påverkar slutledningsförmågan, för att till slut kunna framställa en modell över avvägningen mellan prestandaoptimeringar och slutledningsförmåga.

Vårt examensarbete gick ut på att skapa ett mjukvaruhjälpmedel till utvecklare som vill kunna undersöka olika typer av optimeringar för ett ANN. Programmet kallas *QPANN* och låter en användare enkelt ändra på moduler i ett ANN för att köra enbart heltalsberäkning, m.m. för att sedan se hur slutledningsförmågan förändras. *QPANN* har stor utvecklingspotential, eftersom det går att programmera egna operatorer för att testa hur det påverkar slutledningsförmågan.