FID3025 Konstruktion av effektiv AI-maskinvara genom utforskning av teknologi- och arkitekturdesignrymden 7,5 hp

Architecting Efficient AI Hardware with technology and architectural design space exploration

När kurs inte längre ges har student möjlighet att examineras under ytterligare två läsår.

Fastställande

Skolchef vid EECS-skolan har 2020-06-29 beslutat att fastställa denna kursplan att gälla från och med HT 2020, diarienummer: J-2020-1443.

Betygsskala

P, F

Utbildningsnivå

Forskarnivå

Undervisningsspråk

Undervisningsspråk anges i kurstillfällesinformationen i kurs- och programkatalogen.

Lärandemål

Efter att ha gått denna kurs kommer eleverna att kunna

1. Analysera kraven för ett verkligt maskininlärningsproblem vad gäller lagring, beräkning och effektförbrukning,
2. Ta välgrundade beslut baserat på tillgänglig teknik, arkitektoniska alternativ, exakta uppskattningar av area, prestanda och energi som bäst uppfyller målen för maskininlärningsproblemet,
3. Skapa anpassade AI-lösningar med låg energi som bidrag till en hållbar utveckling,
4. Utvärdera stora forskningstrender och förstå vilka de öppna utmaningar är som samhället fokuserar på.

Kursinnehåll

Kursen består av följande två moduler:

Kravanalys

I den här modulen studerar vi hur man systematiskt extraherar krav på beräkningsoperationer, datatyper, ledningar/kopplingar och datalagring. Dessa krav är logiska till sin natur och oberoende av implementeringsstil. Många exempel från verkligheten kommer att diskuteras under föreläsningar och eleverna får lösa problem för att få praktisk erfarenhet.

Att kunna förstå energikraven är det första steget i att skapa lågenergi-, och därmed hållbara, lösningar.

Konstruera AI-hårdvara och förstå tekniska och arkitektoniska avvägningar

I denna modul studerar vi vad som är de arkitektoniska avvägningarna vid implementering av AI-hårdvara. Vi går in på detaljerna i minneshierarkin och deras teknikalternativ. Minne är den mest dominerande kostnadskomponenten och vi studerar hur man utnyttjar temporär lokalitet för att minimera kostnaden för minneslagring och dataåtkomst.

Bredvid minnet är ledningar/kopplingar den största utmaningen. Ledningar har den värsta skalningsaspekten av tekniken idag. Till exempel, att flytta data med 1 mm på ett chip har jämförbara energikostnader med en singel-precisions flyttalsoperation. Förutom energikostnader spelar ledningar/kopplingar också en stark roll i arkitektoniska beslut. Till exempel är det ett vanligt misstag att öka parallelliteten i beräkningarna utan att samtidigt öka parallelliteten i åtkomst till data. Vi visar hur vi kan göra konstruktioner som möjliggör ökad beräkning med matchande ökning av bandbredden till minnet.

Slutligen studerar vi också vilka alternativ som finns för att implementera de aritmetiska operationerna i Neurala Nätverk. Vi studerar hur man gör avvägningar när det gäller noggrannhet visavi implementeringskostnader med hjälp av en konkret fallstudie från fältet bakterie-genom-igenkänning.

Att känna till dessa arkitektoniska och teknologiska alternativ för att minska energi kommer att bidra till hållbara AI-lösningar.

Särskild behörighet

Registrerad som doktorand.

Examination

- EXA1 - Skriftlig examination, 7,5 hp, betygsskala: P, F

Examinator beslutar, baserat på rekommendation från KTH:s samordnare för funktionsnedsättning, om eventuell anpassad examination för studenter med dokumenterad, varaktig funktionsnedsättning.

Examinator får medge annan examinationsform vid omexamination av enstaka studenter.

Små projekt kommer att definieras med tanke på elevernas eller studentgruppernas forskningsintressen.

Studenterna kommer att tilldelas papper att läsa och presentera för klassen. Deras presentation kommer att användas för att bedöma hur väl de har tagit igenom innehållet i uppsatsen och relaterat till deras problem.

Övriga krav för slutbetyg

För att få godkänt betyg måste eleverna delta i minst 80% av lektionsundervisningen, lämna in alla uppgifter och ge den slutliga presentationen av sitt projekt.

Etiskt förhållningssätt

- Vid grupparbete har alla i gruppen ansvar för gruppens arbete.
- Vid examination ska varje student ärligt redovisa hjälp som erhållits och källor som använts.
- Vid muntlig examination ska varje student kunna redogöra för hela uppgiften och hela lösningen.