INSTITUTIONEN FÖR DATAVETENSKAP | LUNDS TEKNISKA HÖGSKOLA | PRESENTERAD 2024-05-30

EXAMENSARBETE Optimizing Memory Usage of Tensors in Neural Networks
STUDENTER Madeleine Berild, Jakob Sinclair
HANDLEDARE Jonas Skeppstedt (LTH), Kristofer Jonsson (Arm)
EXAMINATOR Flavius Gruian (LTH)

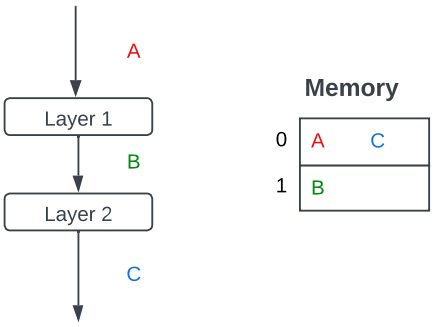
Optimering av neurala nätverk för effektivare AI

POPULÄRVETENSKAPLIG SAMMANFATTNING **Madeleine Berild, Jakob Sinclair**

År 2024 är AI och neurala nätverk mer relevant än någonsin. Användandet av tjänster som Chat-GPT och DALL-E har ökat explosionsartat, och därmed blir det alltmer viktigt att ta hänsyn till hårdvaruresurser för att spara energi. Detta examensarbete undersöker och presenterar flera algoritmer som gör detta möjligt.

Föreställ dig att du äger ett hotell med olika stora rum. Gäster kan boka sin vistelse genom att ange datum och hur många personer som ska bo på rummet. När en gäst checkar in måste rummet värmas upp, och att värma upp rummet kostar mycket pengar. För att spara så mycket uppvärmningskostnader som möjligt är det klokt att ett sällskap flyttar in i ett rum samma dag som ett annat sällskap flyttar ut. Målet blir att varje uppvärmt rum ska användas så mycket som möjligt av tiden, för att undvika att värma upp fler rum än nödvändigt.

Liknelsen med hotellet kan användas för att förstå minnesplanering av data i neurala nätverk. Neurala nätverk kan i vissa fall utnyttja väldigt mycket minne, men allt minne behövs inte användas samtidigt. Viss data behövs bara under kort tid, precis som att gäster på ett hotell oftast bara bor i ett rum under en kort tid. Då är det effektivare att låta annan data utnyttja den platsen efteråt, än att låta minnesplatsen vara outnyttjad och lägga den nya datan på en ny plats. Målet är att minska hur stort minne som behövs, vilket i förlängningen påverkar hur mycket hårdvaruresurser och energi som behövs för att köra de neurala nätverken. Detta är viktigt eftersom det har föruspåtts att AI-tjänster kommer att kräva lika mycket energi som hela Sverige år 2027.



Detta examensarbete presenterar en ny algoritm som kan användas för att undersöka vilken data i ett neuralt nätverk som kan dela plats i minnet på ett grafikkort. Efter det undersöks flera olika algoritmer för att hitta en bra strategi för minnesplanering - dvs, vilken data som ska placeras på vilken minnesadress. Slutsatsen är att det för många neurala nätverk går att spara väldigt mycket minne med dessa algoritmer, ibland uppemot 75%. Många grafikkort har idag stora minnesresurser, men algoritmen kan bli speciellt viktig i de fall då grafikkortet har mycket lite minnesutrymme, som t ex i en mobiltelefon, eller då minnet delas med andra delar av ett system.