

VLSI implementering af neurale netværk

Erik Bruun

Kendetegnende for neurale netværk er det store antal regneoperationer, som forløber i parallel, dvs. samtidigt. Til gengæld er hver operation ret simpel, typisk en multiplikation af et signal med en vægtfaktor (den synaptiske vægt) og addition af resultatet til et fælles signal, eller beregning af en ulineær sammenhæng mellem indgangssignal og udgangssignal for en neuron. Kendetegnende for en traditionel von Neumann computer arkitektur, som f.eks. anvendes i en almindelig PC, er, at beregningsenheden (CPU'en) udfører én operation ad gangen, således at mange operationer nødvendigvis må udføres serielt, dvs. efter hinanden. Det er derfor ikke overraskende, at man i forbindelse med implementering af kunstige neurale netværk undersøger andre beregningsstrukturer end den traditionelle von Neumann arkitektur. Eksempler herpå er computere med en større grad af parallel beregning, f.eks. en 'Connection Machine'. En anden mulighed er at gå helt til de grundlæggende byggeklodser i computerarkitekturen, de VLSI chips, som foretager beregningerne. Herved kan man optimere de enkelte beregningslementer netop med henblik på den type af beregninger, som kræves i et neuralt netværk. Ved Elektronisk Institut, Danmarks Tekniske Højskole, har vi gennem de seneste 3 år arbejdet med konstruktion af specielle VLSI chips til implementering af neurale netværk. Vi har valgt at arbejde med en analog signalrepræsentation i stedet for den digitale signalrepræsentation, der benyttes i traditionelle computere. Årsagen hertil er, at de hyppigst forekommende beregninger i et neuralt netværk (multiplikationer) kan implementeres på et mindre siliciumareal i analog form.

Et neuralt netværk baseret på synapse- og neuron chips.

De grundlæggende beregningsoperationer i et neuralt netværk kan beskrives ved de to ligninger

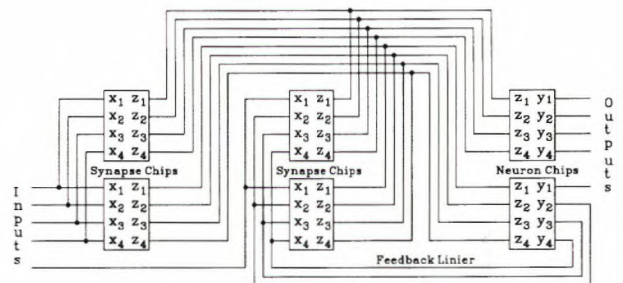
$$y_i = f(z_i) \quad (1)$$

$$z_i = \sum_{j=1}^M w_{ij} x_j \quad (2)$$

Den første ligning udtrykker den ulineære sammenhæng mellem indgangs- og udgangssignaler for en neuron. Funktionen $f(z)$ er ofte en sigmoid (S-formet) funktion, men også andre ulineære funktioner ses anvendt. Den anden ligning udtrykker, at indgangssignalet til en neuron fremkommer som den vægtede sum af et antal (M) signaler, x_j , f.eks. udgangssignaler fra andre neuroner. Summationsvægtene w_{ij} er de synaptiske vægte.

En nærmere betragtning af dette ligningssæt afslører, at x -signalerne er karakteriseret ved at skulle distribueres til et antal synaptiske indgange. Derimod anvendes

hvert z -signal kun ét sted, men til gengæld dannes z -signalerne ved summation af en række produkter, $w_{ij}x_j$, ét for hver synaptisk vægt. Dette fører til den konklusion, at $w_{ij}x_j$ -signalerne og z -signalerne mest hensigtsmæssigt repræsenteres ved analoge signalstrømme, der adderes simpelt hen ved at forbinde signalledningerne sammen til et fælles knudepunkt.



Figur 1. Et neuralt netværk opbygget af neuron- og synapse chips.

Til forskel fra dette repræsenteres x -signalerne bedst ved spændinger, idet en spænding nemt distribueres til mange forskellige indgange. Vi har hermed fået defineret to forskellige moduler, som tilsammen udgør de nødvendige grundelementer til opbygning af neurale netværk med en (i princippet) vilkårlig struktur: et synapse modul og et neuron modul. Figur 1 viser, hvorledes et netværk med 8 neuroner og 5 indgange samt nogle tilbagekoblede signaler kan opbygges ved hjælp af disse to moduler. Der er her antaget, at hver neuron modul indeholder 4 neuroner, og hver synapse modul har 4 indgange og 4 udgange, svarende til 16 synaptiske vægte.

Synapse- og neuron chips.

Synapse modulet kan opbygges som en chip indeholdende dels en (analog) lagring af de synaptiske vægte, dels en analog multiplikator for hver synaptisk vægt. Den analoge multiplikator realiseres i praksis vha. 4 MOS transistorer pr. multiplikator. Lagringen af de analoge vægte er opnået ved at lagre en analog spænding repræsenterende vægten på en kondensator. Da en sådan kondensator har en vis lækstrøm, vil den imidlertid efterhånden aflades til en bestemt, fast værdi og således miste informationen. Det er derfor nødvendigt med en regelmæssig 'refresh' af de synaptiske vægte. Dette sker ved at adressere den ønskede kondensator via nogle MOS transistorer, der fungerer som elektroniske omskiftere, og påtrykke en analog spænding svarende til den ønskede vægt. Den analoge spænding

fås fra et eksternt, digitalt vægtlager med en tilhørende D/A konverter. Et blokdiagram af systemet er vist i figur 2. Dette system indebærer ganske vist en 'dobbelt' lagring af de synaptiske vægte, dels i det eksterne, digitale vægtlager, dels i det interne, analoge arbejdslager på synapse chippen. Til gengæld kan der til synapsechippen anvendes en halvlederproces, der er optimeret med hensyn til de analoge kredsløbsfunktioner, mens der til det digitale vægtlager kan anvendes en optimal teknologi til digitale hukommelseskredsløb.

Neuron chippen realiseres ved hjælp af et forholdsvis simpelt kredsløb, som anvender en bipolar transistor til opnåelse af en hyperbolsk tangens overføringsfunktion.

De udviklede chips har følgende væsentlige karakteristika:

	Synapse chip	Neuron chip
Areal pr. synapse/neuron	0.015mm ²	0.40mm ²
Forsinkelse fra indgang til udgang	2.5µs	0.5µs
Nøjagtighed	3%	2%
Opløsning	0.2%	—

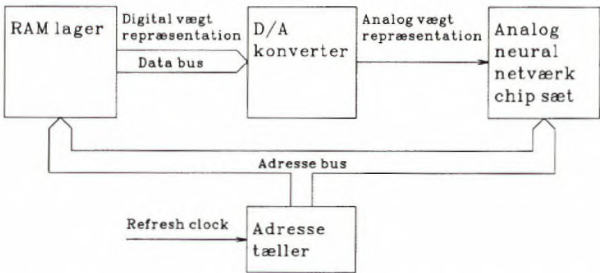
På basis heraf kan det estimeres, at der med chipstørrelser på mindre end 40mm² vil kunne udvikles et chip sæt med 50 neuroner og 50×50 synaptiske vægte pr. chip. Herved er det realistisk at tale om systemer med flere hundrede neuroner og en beregningskapacitet i størrelsesordenen 10¹⁰ conn./sec.

Indlæring.

En ting er, at neurale netværk vha. VLSI teknologi kan konstrueres til at give en meget hurtig beregning i 'recall' fasen af netværket, dvs. når netværket først er oplært – alle de synaptiske vægte er fastlagt. Noget andet er selve beregningen af de synaptiske vægte gennem en passende indlæringsalgoritme. Denne side af sagen er typisk endnu mere beregningskrævende end selve 'recall' fasen, både fordi indlæringsalgoritmen er vanskeligere og fordi beregningerne skal gennemføres på et stort antal eksempler. Derfor er en VLSI integration af indlæringsalgoritmen mindst lige så påkrævet som integration af det neurale netværks 'recall' algoritme.

Ved Elektronisk Institut arbejder vi også med udvikling af chips til indlæring. I dette arbejde foretages først en analyse og dekomponering af beregningsalgoritmen til indlæring og dernæst en direkte VLSI implementering af de enkelte elementer i algoritmen. Vi har eftervist, at en af de hyppigst anvendte indlæringsalgoritmer (backpropagation) kan implementeres i VLSI på en sådan måde, at

de beregningslementer (neuron- og synapse chips), der anvendes i 'recall' fasen, med beskedne modifikationer kan anvendes i indlæringsfasen, således at der næsten ikke kræves ekstra siliciumareal til indlæringsalgoritmen. Vi har endvidere vist, hvorledes mere komplicerede indlæringsalgoritmer kan dekomponeres på en sådan måde, at der opnås et kompromis mellem kompleksitet og beregningshastighed, derved at dele af algoritmen udføres i parallel, mens andre dele udføres serielt.



Figur 2. Vægtlagringssystemet i et neuralt netværk med en dynamisk analog hukommelse.

Fremtiden

De eksperimentelle systemer, vi indtil videre har lavet, er ganske små. Vores første generation af neuron chips og synapse chips indeholder hhv. 4 neuroner og 4×4 synaptiske vægte. Forsiden på dette nummer af KVANT viser et foto af synapse chippen. Disse chips er nu indbygget i en PC og kan anvendes til demonstration af meget simple netværksopgaver. Fremover vil vi koncentrere os om meget større systemer, hvor man virkelig udnytter VLSI teknologiens fordele til frembringelse af store systemer med et forhold mellem pris og ydeevne, som vanskeligt kan opnås på anden vis. Vores mål er at opbygge et stort netværk med oplæring i hardware, der f.eks. kan benyttes til forudsigelse af splice sites i pre-mRNA molekyler.



Erik Bruun Professor i analog elektronik ved Elektronisk Institut, Danmarks Tekniske Højskole, siden 1989. Har i 20 år arbejdet med udvikling af LSI og VLSI kredsløb. Var fra 1974 til 1980 ansat ved Laboratoriet for Halvlederteknik, DTH, fra 1980 til 1984 hos Christian Rovsing A/S, og fra 1984 til 1989 hos Danmos Microsystems ApS.