확장성과 범용성을 갖는 신경망 프로세서

한 건 희 / 연세대학교 전기전자공학과

지능 프로세서의 필요성

향후 21세기는 여러가지 장치들과 인간과의 보다 자연스러운 인터페이스 대한 요구가 증가될 것이며 대부분의 일상용품(컴퓨터, TV, 인터넷, 각종 가전제품, 자동차, 휴대용 정보 기기 등)에 이러한 기능이 추가될 것으로 기대된다. 보다 인간 친화적인 인터페이스를 위해서는 주어진 상황에서 최적화가 이루어지는 학습기능이 필수적이게 될 것이다. 특히 범용성이 큰 대형 시스템의 환경 설정이 무척 복잡하게되고 일반 사용자가 자신이 원하는 형태로 시스템을 설정하기가 어려워 인간친화적 기법으로서 학습기능이 내장될 것으로 예상된다. 이러한 지능의 구현은 〈그림 1〉과 같이 현재 디지탈 컴퓨터의 계산량과 속도를 능가하는 시스템이 요구된다.

마이크로 프로세서의 계산속도는 〈그림 2〉와 같이 대략 매 5년마다 10배의 향상을 보이고 있으며 메모리 또한 비슷한 경향을 나타내고 있다. 이와 같은 기술 동향을 고려할 때 특별한 기술적인 장벽이 없다면 대략 40년 후에는 하드웨어적으로는 인간의 뇌에 상응하는 computing power와 memory가 오늘날의 PC와 같은 시스템 크기와 가격으로 구현 될 수 있을 것으로 예측할 수 있다.

범용성을 중심으로 하는 CPU는 특정 응용분야에서는 지나치게 복잡성이 커서 효율적이지 못하거나 또는 계산속도의 한계로 실시간 응용이 어렵다. 범용 CPU로서는 얻기 어려운 계산속도를 얻기 위하여 DSP가 개발되었으며, 특히 DSP는 고속계산에 적합한 구조를 가지고 있어 같은 공정 기술을 사용한 CPU에 비하여 계산속도가 매우 빨라 음성 또는 연상신호처리 뿐만 아니라 고속 정보전송 및 저장을 가능하게 하였다. 하드웨어 구현 방법에 있어 서는 ASIC 칩으로 구현하는 경우 개발 대상응용 분야에서는 최적의 성능 및 효율을 제공하나 다양한 응용분야에 적용하기 어렵다는 단점이 있다. 이와 반대로 FPGA의 경우는 임의의 시스템을 손쉽게 chip으로 구현이 가능하나 동작속도가 느리고 구현 가능한 시스템의 크기 도한 제한적이다. 이러한 프로세서 구조와 구현방법의 발달에도 불구하고 큰 용량과 빠른 속도 범용성을 모두 요구하는 지능 모듈을 구현하기에는 부족한 실정이다.

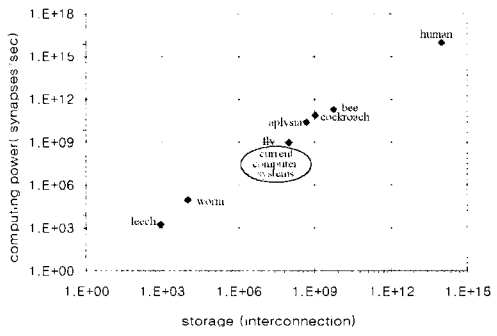


그림 1 컴퓨터의 계산 능력

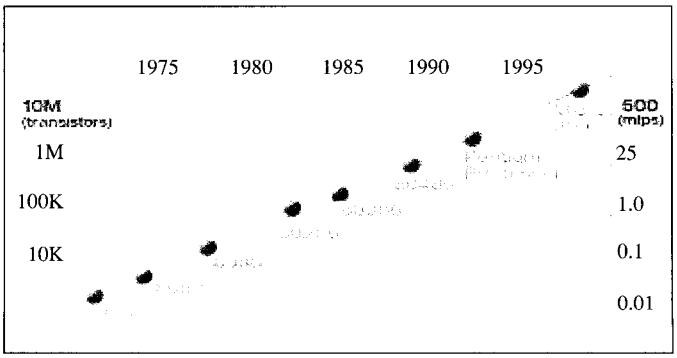


그림 2 Microprocessor 기술 동향

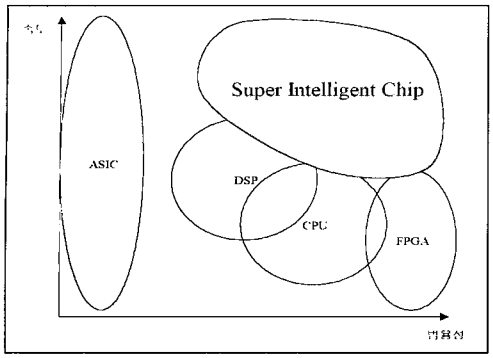


그림 3 프로세서에 따른 범용성과 속도의관계

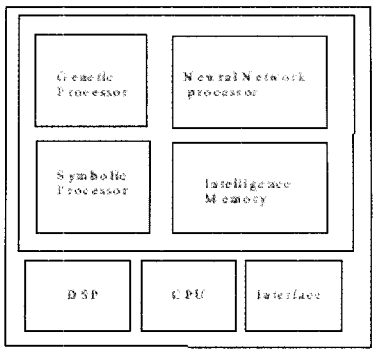


그림 4 통합 지능모듈

지능프로세서와 신경망 프로세서 코어

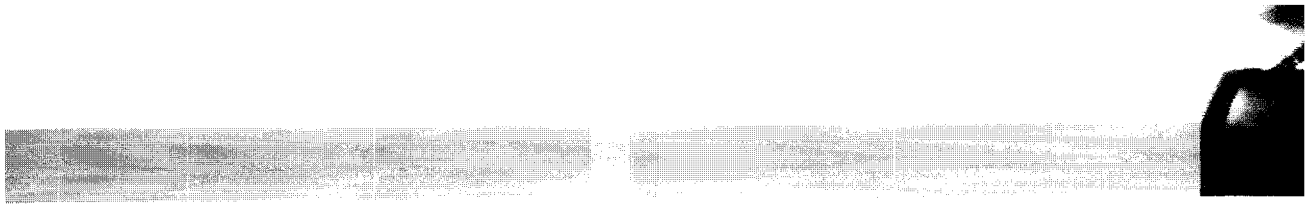
지능은 다양하고 복잡한 연산에 의하여 이루어지므로 어떠한 단일한 구조만으로는 지능의 구현이 불가능하다. 실질적으로 인간의 뇌에서도 여러 지능의 요소(기억, 계산, 운동, 감정등)가 뇌의 특정 부분과 각각 연관되어 있으며 이는 각각의 영역의 특화 되어있음을 의미한다. 지금까지 알려진 지능 알고리즘으로는 신경망 알고리즘, 진화 알고리즘, 퍼지로직, symbolic processing 등이 있으며 어느 특정 알고리즘만으로는 고도의 지능을 구현하기 어렵다. 그러므로 지능 모듈은 〈그림 4〉 와 같이 다양한 지능 프로세서와 이들을 제어하기 위한 CPU, 다양한 신호를 처리하기 위한 DSP 및 인터페이스 모듈들이 통합되어짐으로서 보다 복잡한 지능을 구현 할 수 있을 것으로 예상된다.

이러한 지능 알고리즘중 신경망 알고리즘은 가장 널리 사용되는 알고리즘으로 수치 해석적 최적화 알고리즘의 한 부류로 생각할 수 있으며 대용량의 병렬연산을 요구한다. 다양한 응용분야가 보고되고 있으나 계산량의 한계로 실용성 있는 시스템의 구현이 어려운 경우가 많이 있다. 또한 신경망 알고리즘에도 여러 다른 알고리즘과 응용 분야 등이 제시되어 이러한 서로 다른 알고리즘을 처리할 수 있는 프로세서가 요구된다.

신경망 프로세서를 포함한 여러 지능 모듈간의 역할분담 및 상호작용에 의하여 보다 고차원적인 학습, 판단, 추론, 인식능력을 갖춘 지능 모듈이 구현될 수 있으므로 신경망 프로세서는 지능 모듈의 필수적인 코어 프로세서라 할 수 있다.

신경망 프로세서 기술동향

신경망 알고리즘에 대한 연구가 활발히 진행되면서 다양한 응용분야에 적용하기 위하여 고속 계산이 가능한 신경망 하드웨어 와 신경망 analog VLSI, 신경망 프로세서에 대한 연구가 미국 NSF, NASA, Office of Naval Research등의 지원 아래 활발하게 이루어졌으며 특히 디지탈 시스템의 속도가 한계에 부딪힐 것으로 예상된 80년대 중반부터 analog 신경망 VLSI에 대한 연구가 매우 활발하게 이루어졌다. 그러나 90년대 초반부터 반도체 공정 기술의 발달과 디지탈 시스템 아키텍춰의 발달로 디지탈 시스템의 속도와 용량이 크게 향상되면서 analog 신경망 VLSI에 대한 연구는 점차 쇠퇴하게 되었다. Analog 신경망 VLSI는 고속 병렬처리를 매우 적은 실리콘 면적과 전력소모로 구현할 수 있다는 장점을 가지고 있으나 개발된 chip의 범용성이 크게 제한되어 있으며 수율이 낮아 상업적 제품으로 성공하지 못하였다. 90년대 중반부터 후반까지는 신경망 VLSI에 대한 연구가 다소 쇠퇴하였으나 미국에서는 국방과 항공우주관련 연구소와 산업체에서 꾸준한 투자가 진행되었다. 90년대 후반에 들어서 디지탈 반도체 기술의 비약적인 발전에 힘입어 신경망 VLSI에 대한 연구가 다시 활성화되기 시작하였다. 90년대 초와 중반에는 수많은 신경망 software들이 계산속도의 한계로 인하여 실용성을 얻지 못하였으나 90년도 후반부터 DSP와 같은 고속 대용량 processor 개발에 힘입어 실용성을 얻게 되었다. 그러나 많은 응용분야에서 아직도 요구되는 계산 속도를 얻지 못하고 있으며 DSP를 사용하여 신경망 알고리즘을 처리하는 경우 DSP칩의 크기와 전력 소모등을 고려할 때 그 효율이 매우 낮아 이동성이 요구되는 시스템, 소형 시스템 또는 일상 가전 용품에는 적용되지 못하고 있는 실정이다. 이러한 시장의 요구에 의해 90년대 후반부터 상용 디지탈 신경망 프로세서들이 속속 개발되었다. 예를 들어 Accurate Automation Corporation이라는 회사는 8,000개의 neuron이 완전하게 서로 연결될 수 있으며 초당 140M synapse를 처리할 수



있는 신경망 프로세서를 개발하였다. 이 프로세서와 DSP가 통합된 시스템을 판매하고 있으며 이를 NASA와 미 공군의 초소형 무인 원격조정 초음속 비행체 개발 계획(LoFLYTE)에 사용되어 성공적으로 실험비행이 이루어졌다. AMS와 협력관계를 갖는 Italy의 Neuricam이라는 회사는 32개의 neuron과 각 neuron 당 256 weights 를 갖는 저가형 신경망 프로세서를 개발 $100이하의 가격으로 판매하고 있으며 1~1024개의 신경망 프로세서와 Pentium급 CPU가 통합된 시스템을 $30,000이상의 가격으로 판매하고 있다. 러시아의 Module이라는 회사는 32bit RISC core와 64-bit Vector processor core가 결합된 구조를 단일 chip에 구현하여 14,400 MMAC(Million Multiplication and Accumulation per second)의 계산속도를 가지며 16Gb의 memory를 control할 수있을 뿐만 아니라 DSP와도 손쉽게 통합시스템을 꾸밀 수 있는 신경망 프로세서를 개발 판매중이다. 액세온(Axeon)이라는 벤쳐 기업에서는 256개의 간단한 8비트 RISC 프로세서들을 한 개의 칩에 구현 초당 2.4G sysnapse를 계산할 수 있는 processor를 개발 연내 3세대 무선통신이나 관성항해 및 자동화된 이미지 해석 등과 같이 다양한 응용분야에 적용할 계획이다.

신경망 프로세서 구조

신경망 알고리즘은 일반적으로 training과 recall의 두 부분으로 나뉘어 지며 가장 흔히 사용되는 구조가 Multi-layer Perceptron (MLP)이며, 학습 알고리즘으로는 Back Propagation(BP)이 널리 사용된다. 그러나 다양한 알고리즘이 응용분야에 따라 사용되며, 그 구조가 다르므로 reconfigurable하거나 프로그램이 가능한 구조가 요구된다.

신경망 알고리즘은 기본적으로 다량의 곱셈과 덧셈을 요구하는 알고리즘으로 이를 구현하기 위한 신경망 프로세서 구조는 크게 두 가지 접근 방법으로 분류될 수 있다. 첫째는 고성능 vector multiplier와 cache memory를 사용하는 방법으로 범용성이 뛰어나다는 장점이 있으나 신경망 알고리즘의 특성상 메모리 접근이 많아 여기에서 병목 현상이 일어날 수 있다는 단점이 있다. 이러한 구조는 신경망 프로세서보다는 DSP에 가까운 구조라 할 수 있다. 또 다른 방법으로는 신경망 알고리즘에서 사용되는 연산이 주로 곱과 합의 단순 연산인 점을 고려하여 비교적 단순한 구조의 processing element(PE)를 수십 또는 수백개 구현하고 연산이 병렬적으로 수행되도록 하는 구조로 메모리 접근 시 병목 현상을 최소화하기 위하여 local memory를 각 PE에 분산 배열함으로서 1개의 PE와 local momory를 1개의 processing unit(PU)로 구현하여 이들을 병렬 연결하는 구조이다. 병렬구조는 계산속도가 빠르다는 장점이 있으나 병렬성이 적은 응용 분야에서는 많은 PU가 대기 상태로 있게 되어 계산 효율이 떨어진다는 단점이 있다.

이러한 병렬처리 시스템[2]의 구조는 Single Instruction Multiple Data(SIMD) 구조와 Multiple Instruction Multiple Data(MIMD)구조로 나뉠 수 있으며 신경망 알고리즘은 동일한 계산이 반복되므로 SIMD구조가 적합하다 하겠다. 특히 신경망 알고리즘은 data양이 많으므로 memory access에서의 overhead를 최소화 할 수 있는 구조가 요구된다.

신경망 프로세서 구조 개발에 있어서 중요한 고려 사항이 범용성과 확장성의 확보라할 수 있다. 90년대 개발된 신경망 프로세서가 시장 진입에 실패한 가장 큰 원인이 범용성과 확장성 결여라 할 수 있다. 응용분야에 따라 신경망의 크기가 매우 다르며 알고리즘 또한 다양하여 한정된 몇몇 알고리즘이 지원되는 대형 프로세서는 시장성을 확보하기 어렵다. 그러므로 비교적 적은 개수의 PU를 단일 칩으로 구현하고 다수개의 칩을 손쉽게 연결하여 보다 큰 시스템을 구현할 수 있어야만 다양한 응용 분야에 사용될 수 있다. 또한 여러 신경망 알고리즘뿐만 아니라 서로 다른 알고리즘의 조합 또는 각종 신호 및 영상처리가 가능하도록 software에 의하여 프로그래밍이 가능한 구조를 갖추어야만 시장성을 확보할 수 있다.

이러한 대규모의 신경망 프로세서는 반도체 공정의 발달로 이러한 큰 시스템을 단일 칩으로 구현이 가능하게 될 것이며 특히 이러한 복잡한 시스템의 설계 및 제조는 반도체 산업의 자연스러운 발전 방향인 System-On-a-Chip(SOC) 기술 발달의 직접적인 혜택을 보게될 것이다. 신경망 프로세서의 구현에 있어서 가장 큰 칩 면적을 차지하는 것이 메모리로 신경망 프로세서는 다량의 memory access가 요구되며 외부 메모리를 사용하는 경우 Memory access 시간과 이를 위한 전력소모가 매우 클 것으로 예상된다. 향후 대용량 로직이 단일 칩에 구현되는 Memory Merged Logic 공정기술의 발달로 대용량의 메모리를 프로세서와 동일 한 chip에 구현할 수 있을 것으로 기대된다.