論文2003-40CI-6-7

대규모 확장이 가능한 범용 신경망 연산기 : ERNIE

(Expansible & Reconfigurable Neuro Informatics Engine : ERNIE)

金 榮 柱 *, 童 聖 秀 **, 李 鍾 浩 ***

(Young-Joo Kim, Sung-Soo Dong, and Chong-Ho Lee)

요 약

범용 신경망 연산기를 디지털 회로로 구현함에 있어 가장 까다로운 문제들 중 하나는 시냅스의 확장과 해당 네트워크에 맞게 뉴런들을 재배치하는 재구성 문제일 것이다. 본 논문에서는 이러한 문제들을 해결하기 위한 새로운 하드웨어 구조를 제안한다. 제안된 구조는 시냅스의 확장과 네트워크 구조의 변경을 위해 오리지날 디자인의 변경이 필요치 않으며, 모듈러 프로세싱 유니트의 확장을 통한 뉴런의 개수 및 레이어의 확장이 가능 하다. 이 구조의 범용성 및 확장성에 대한 검증을 위해 다양한 종류의 다층 퍼셉트론 및 코호넨 네트워크를 구성하여 HDL 시뮬레이터를 통한 결과와 C 언어로 작성된 소프트웨어 시뮬레이터 결과를 비교 하였으며 그 결과 성능이 거의 일치함을 확인하였다.

Abstract

Difficult problems in implementing digital neural network hardware are the extension of synapses and the programmability for relocating neurons. In this paper, the structure of a new hardware is proposed for solving these problems. Our structure based on traditional SIMD can be dynamically and easily reconfigured connections of network without synthesizing and mapping original design for each use. Using additional modular processing unit the numbers of neurons and synapses increase. To show the extensibility of our structure, various models of neural networks : multi-layer perceptrons and Kohonen network are formed and tested. The performance comparison with software simulation shows its superiority in the aspects of performance and flexibility.

Keywords : Neural network, Digital hardware, SIMD, Modular structure.

* 學生會員, *** 正會員, 仁荷大學校 情報通信工學科
(Dept. of Information technology & Telecommunication, Inha University)
** 正會員, 龍仁松潭大學 디지털電子情報科 專任講師
(Dept. of Digital Electronics & Information, Yong-in Songdam College)

※ This research was supported as a Brain Neuroinformatics Research Program sponsored by Korean Ministry of Science and Technology.
接受日字:2003年9月1日, 수정완료일:2003年11月7日

Ⅰ. 서 론

신경회로망은 뇌의 정보처리 방식을 모델화 한 것으로써 복잡한 비선형 시스템의 제어에 적합하며 연상, 추론, 인식 등 기존의 컴퓨터가 해결하기 어려운 응용분야에서 많이 사용되고 있는 알고리즘이다. 이러한 신경회로망은 병렬, 분산 처리를 본질적인 특징으로 가지기 때문에 소프트웨어 시뮬레이션만으로는 한계를 가지게 되어 1958년 F. Rosenblatt의 MARK I Perceptron 이후로 신경회로망을 하드웨어로 구현하려는 노력이 계속되어져 왔다[1]. 그중에서도 근래에 들어

서는 디지털 전자회로기술의 발전에 힘입어 아날로그 신경회로망보다는 디지털 신경회로망의 구현으로 더욱 더 많은 연구가 있어왔다.

디지털 신경회로망을 구현하기 위한 방법은 크게 두 가지로 분류할 수 있다. 첫째는 DSP를 이용한 범용 프로세서 기반의 신경망이며, 둘째는 시냅스, 뉴런 등의 회로를 병렬구성 하여 신경망 연산 전용 칩을 ASIC으로 구현하는 것이다[2]. 전자의 경우 후자에 비해 자유도가 높으며 구현이 쉽다는 장점이 있으나 뉴런의 선형적인 증가에 대해서 시냅스가 지수적으로 증가하는 데에 따르는 연산 시간 및 면적의 증가를 해결하기 어려우며 범용 연산기의 사용은 하드웨어 구현시 상대적으로 큰 면적을 차지하게 된다. 특히 디지털 회로에서의 곱셈기는 아날로그에 비하여 큰 면적을 차지하게 되는데 이를 극복하기 위하여 곱셈기가 없는 디지털 신경망[3], 또는 곱셈기의 크기를 줄이기 위한 시리얼 데이터 방식의 디지털 신경망[4]과 같은 연구가 있었다. 후자의 경우 자유도 문제에 있어서는 취약하지만 신경망 연산을 위한 전용회로를 사용하기 때문에 전자의 경우보다 면적문제에 있어 많은 이점을 가진다. 그러나 이 경우에도 역시 곱셈기의 증가 문제는 회로의 집적도를 크게 떨어뜨리는 요인으로 작용한다.

하드웨어로 구현된 신경망 연산기가 처리속도의 측면에서 소프트웨어로 구현된 신경회로망 시뮬레이터에 비해 월등한 이점을 갖는다고 해도 정형화된 가중치 혹은 특정한 형태로 고정된 네트워크의 기능만을 수행한다면 다양한 공학적인 분야에서 그 이용가치가 떨어지게 되므로 특정 수준 이상의 범용성은 디지털 신경망 연산기의 필수조건이라 할 수 있다. 이를 위해서 다양한 연구가 있어왔으나 주로 범용 프로세서를 이용하여 신경망 연산기를 재프로그래밍 하는 방법과[5] 마치 FPGA처럼 간단한 재구성 비트를 이용하여 기본 PE(Processing Element)의 역할과 버스구조를 변경함으로써 원하는 형태의 신경회로망을 구성하는 재구성형 하드웨어 구조 등이 제시되어 왔다[6]. 범용 프로세서를 이용할 경우 우수한 자유도를 가지지만 처리속도와 면적면에서 취약점을 드러내며, 재구성형 하드웨어에 기반한 신경망 연산기는 제한된 자유도를 갖지만, 빠른 처리속도와 적은 면적에 구현가능하며 단일 칩 제작이 용이하다는 것이 장점이다.

B. Pino는 각각의 PE의 시스템 버스를 인접한 PE 사이에만 연결함으로써 PE의 직렬 확장을 통해 네트워크의 크기를 확장하는 방법을 소개하였다[7]. 이러한 구조를 통해 PE의 한계에 제한을 받지 않으면서 원하는 규모의 네트워크를 구성할 수 있지만 PE사이의 버스를 통해서만 데이터의 전송이 이루어지기 때문에 병렬성의 측면에 있어서 바람직하지 못한 결과를 가져오게 된다. 또한 B. Girau는 FPGA의 재구성 원리를 이용한 FPNA(Field Programmable Neural Arrays)라는 개념을 통해서 뉴런의 개수를 확장하는 방법을 도입하였지만 이 방법은 각각의 PE가 가지는 한계 이상의 시냅스를 처리할 수 없다는 문제점을 가지고 있다[6].

위에서 제기한바와 같이 시냅스의 지수적 증가에 따른 면적의 증가, 신경회로망의 재구성, 그리고 네트워크의 확장에 관한 문제점들을 효과적으로 극복하기 위하여 본 논문은 SIMD(Single Instruction Multiple Data) 구조 및 마스터-슬레이브(Master-Slave)구조를 응용한 새로운 형태의 모듈러 신경망 구조를 제안하였다.

II. SIMD(Single Instruction Multiple Data)

하나의 명령으로 여러 유니트들이 동시에 동작하는 SIMD 구조는 많은 양의 로컬 데이터를 병렬적으로 처리해야하는 영상 처리나 신경회로망과 같은 어플리케이션을 구현하는데 있어 매우 유용하다. 각각의 PE는 최소한의 기능만을 수행하는 간단한 구조여야 하며 자신만의 로컬 메모리를 갖고 있다. 또한 모든 PE들은 상호간의 네트워크를 통해 연결되어 있으며 각각의 PE들은 배열(array)을 이루고 있다. 각 PE들은 특정 명령을 동시에 받게 되고 자신에게 속해 있는 지역 메모리의 데이터를 해당 명령에 의해 동일하게 처리한다. 필요한 경우 서로간의 연결망을 통해서 처리된 데이터를

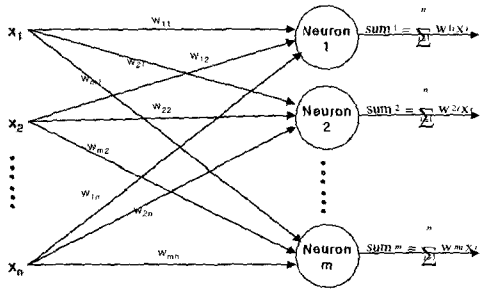


그림 1. 일반적인 신경회로망
Fig. 1. General Neural Network architecture.

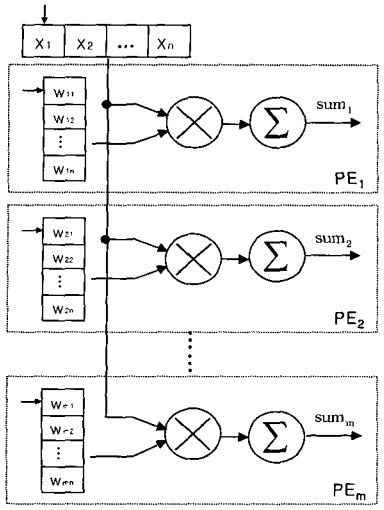


그림 2. SIMD 구조를 이용한 신경회로망
Fig. 2. Neural Network using SIMD architecture.

주고 받는 것이 가능하다[8]. 이러한 SIMD 구조를 신경회로망에 적용할 경우 PE는 뉴런의 역할을 하게 되고 각각의 PE에 속하는 지역 메모리는 시냅스의 가중치를 저장하는 역할을 하게 된다. <그림 1>은 일반적인 신경회로망 모델을 나타내며 <그림 2>는 SIMD 구조를 이용한 신경회로망의 대표적인 예를 보여준다.

이상과 같이 대개의 SIMD 구조에 있어서 모든 PE들은 하나의 지시에 대하여 각각의 데이터를 동일한 방식으로 처리한다. 그러나 본 논문에서는 이를 응용하여 각 PE의 설정에 따라 서로 다른 동작을 수행함으로써 신경회로망의 재구성 및 시냅스 확장을 가능케 하는 하드웨어 구조를 제안한다.

Ⅲ. ERNIE(Expansible & Reconfigurable Neuro Informatics Engine)

1. MPU(Modular Processing Unit)

일반적인 하드웨어 신경망 연산기에서 뉴런의 시냅스 개수는 PE가 갖고 있는 내부 메모리의 크기에 의해서 결정된다. 따라서 뉴런의 역할을 하는 PE의 개수를 확장한다고 해도 내부 메모리 용량을 초과하는 입력은 받아들일 수가 없기 때문에 뉴런의 개수는 늘어나지만 뉴런 당 입력의 개수는 제한을 받게 된다. 본 논문은 기존의 확장 가능한 신경망 연산기 구조가 가지고 있는 이러한 결점을 보완하기 위하여 뉴런의 기능을 수행하는 PE를 내부 연결 버스를 통해 묶음으로써 시냅스 개수의 확장이 가능하도록 하는 형태의 하드웨어 구조를 제안하였다. 제안된 구조를 이용하면 시냅스의 확장은 물론 네트워크를 구성하는 가장 큰 기본 모듈인 프로세싱 유니트(Processing Unit)의 연결을 통하여 뉴런 및 레이어의 확장까지 가능하다. 다수의 PE로 구성되어 있으며 그 자체로서 한 층의 레이어를 구현할 수 있는 이 모듈을 MPU(Modular Processing Unit)라고 명명하였다.

MPU는 두 가지의 PE로 구성되고 설정모드에서 결정되는 상태에 따라 각기 다른 동작을 수행하며 주요 동작은 <표 1>과 같다.

표 1. MPU를 구성하는 두 가지 PE
Table. 1. Two PEs composing MPU.

PE	기 능	내 용	특 징
SPE (Synapse Processing Element)	서밍 노드 (summing node)[9]	입력과 가중치의 MAC 연산	기본 신경망 연산 소자로써 상호 연결에 의하여 시냅스 확장 구현
	코호넨 노드	입력과 가중치의 거리 (distance) 연산	
LPE (Layer Processing Element)	활성화 함수 LUT (Look Up Table)	SPE의 출력을 LUT의 주소값으로 이용	모든 출력을 다른 MPU의 입력으로 사용함으로써 MPU간의 모듈러 확장 구현
	승자 뉴런의 결정	가장 작은 출력을 내는 SPE의 인덱스 출력	

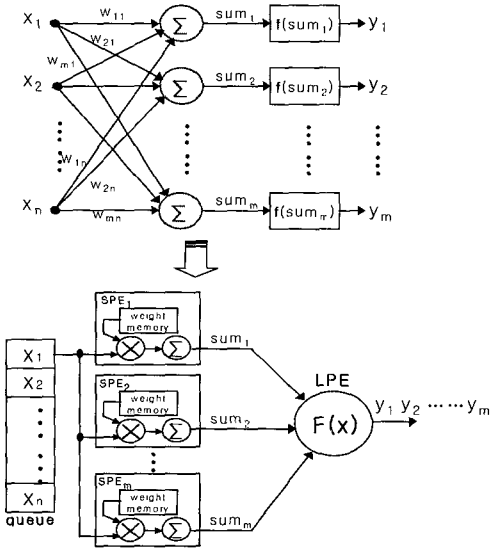


그림 3. 퍼셉트론과 상응하는 MPU의 개략도
Fig. 3. Schematic diagram of MPU corresponding to perceptron.

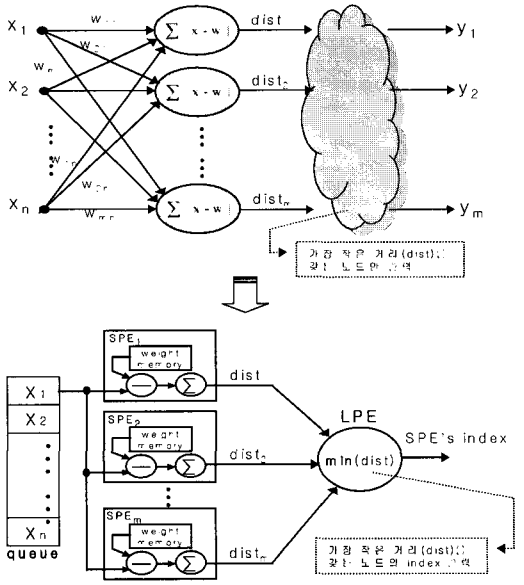


그림 4. 코호넨 네트워크와 상응하는 MPU의 개략도
Fig. 4. Schematic diagram of MPU corresponding to Kohonen network.

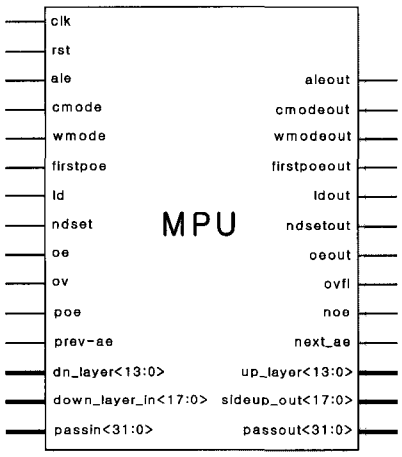


그림 5. MPU의 입/출력 신호
Fig. 5. Input/output signal of MPU.

MPU는 다수의 SPE와 LPE 1개로 구성이 가능하며 구조의 특성상 하나의 MPU를 구성하는 SPE의 개수에 대한 제한은 없다. 본 논문은 SPE 24개와 LPE 1개로 이루어진 MPU를 기준으로 하였으며 실험 또한 이에 맞추어 진행되었음을 밝혀둔다. MPU는 퍼셉트론과 코호넨 네트워크를 구성할 수 있는 최소 단위의 모듈이다. <그림 3>과 <그림 4>는 퍼셉트론과 코호넨 네트워크의 동작을 MPU의 각 PE들이 어떻게 구현하는지를 개념적으로 보여주고 있다.

MPU는 모듈러 확장이 가능하도록 설계되었다. 따라서 MPU의 출력은 또 다른 MPU의 입력으로 사용되며 이러한 모듈러 확장은 MPU들로 구성되는 칩간 확장으로까지 이어진다. <그림 5>를 통하여 MPU의 클럭(clk)과 초기화(rst) 신호를 제외한 모든 입·출력 형식이 동일함을 알 수 있다.

MPU의 입력값은 SPE를 거쳐 병렬적으로 연산이 수행되고 이때의 결과값과 여러 컨트롤 신호들은 LPE를 통하여 다음 MPU의 입력으로 들어간다. <그림 6>은 SPE 24개와 LPE 1개가 MPU를 구성하고 있는 블록 다이어그램이며 이들 PE간에 컨트롤 신호와 데이터가 어떻게 흘러가는지를 보여주고 있다.

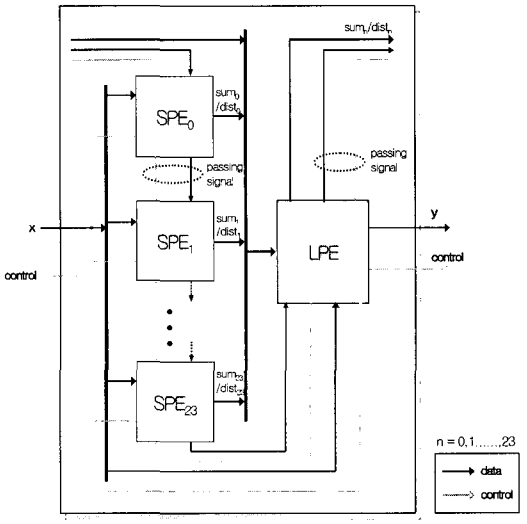


그림 6. MPU의 블록 다이어그램
Fig. 6. Block diagram of MPU.

제안된 구조는 네트워크의 확장을 위한 두 가지 특성을 갖고 있다. 첫 번째는 MPU간의 확장이며 두 번째는 SPE간의 확장이다. MPU간의 확장을 통해서 네트워크를 구성하는 뉴런의 개수 및 레이어의 개수를 확장할 수 있고, SPE간의 내부 연결 버스를 사용하여 하나의 SPE에서 누적된 시냅스와 입력의 곱을 인접한 SPE로 전달하는 방법으로 시냅스를 확장하는 것이 가능하다. 이와 같은 방법을 통하여 PE가 갖고 있는 지역 메모리의 용량을 초과하는 시냅스를 구현하는 문제를 해결할 수 있다. 이러한 형태의 '시냅스 확장'은 둘 혹은 그 이상의 SPE로 구현하는 것이 가능하며, MPU 내부에서만 국한되는 것이 아니라 외부의 MPU에 속해 있는 SPE와도 연결이 가능하기 때문에 대용량의 시냅

스를 요구하는 신경망을 매우 융통성 있게 구현할 수 있다. <그림 7> 및 <그림 8>은 이와 같은 확장이 어떻게 이루어지는가를 보여주고 있다. <그림 7>은 두 개의 SPE를 내부 버스로 연결하여 하나의 뉴런으로 구성한 모습이며 <그림 8>은 두 개의 MPU를 연결하여 레이어를 확장한 모습이다. 이처럼 MPU 2개를 이용하여 네트워크를 구성한 경우 2개의 MPU를 모두 출력층으로 설정하여 최대 48개의 출력 뉴런을 갖는 단층 퍼셉트론을 구현하거나 MPU_0를 은닉층으로 설정하고 MPU_1을 출력층으로 설정하여 각 층마다 24개의 뉴런을 갖는 다층 퍼셉트론을 구성 할 수 있다.

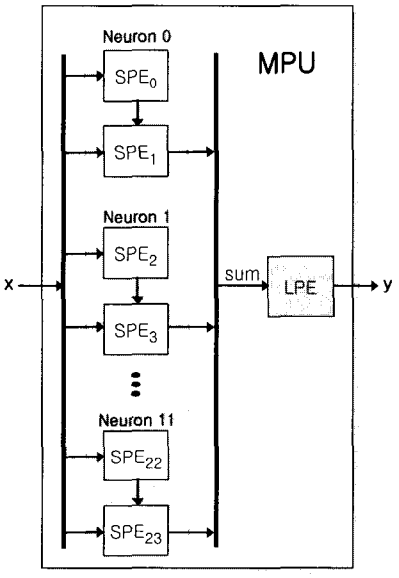


그림 7. 내부연결버스를 통한 시냅스 확장
Fig. 7. Expansion of synapses using internal bus.

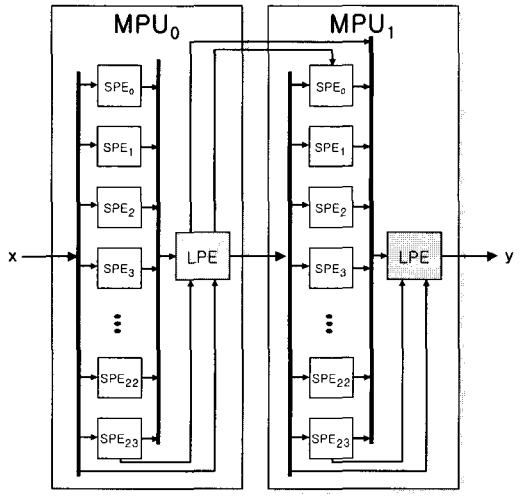


그림 8. MPU의 연결을 통한 레이어의 확장
Fig. 8. Expansion of layers using MPU connections.

모든 유니트 사이(SPE-SPE, SPE-LPE, MPU-MPU)에는 두 가지 종류의 마스터-슬레이브 형태의 버스가 존재한다. 첫 번째는 컨트롤 신호의 전달을 위한 것이고 두 번째는 여러 가지 형태의 확장을 위한 것으로서 특정한 경우에만 활성화 되며 네트워크를 구성하는 모든 기본 모듈을 유기적으로 연결해주는 역할을 한다. 이를 <표 2>와 같이 정리하였다.

표 2. 확장을 위한 마스터-슬레이브 버스의 기능
Table. 2. Functions of master-slave buses for expansion.

연결 형태	기 능	내 용
SPE-SPE	시냅스의 확장	마스터 SPE의 누적합을 슬레이브 SPE로 전달
SPE-LPE	외부 MPU에 포함된 SPE와의 시냅스 확장	마스터 SPE의 누적합을 LPE를 통해 슬레이브 SPE로 전달
MPU-MPU	뉴런 및 레이어의 확장	바이패스(bypass)

MPU를 구성하는 모든 PE는 동작 초기에 설정 모드를 통하여 그 역할이 결정된다. 이러한 과정을 거침으로써 매우 다양한 형태의 신경망 연산기를 구성하는 것이 가능하다. 따라서 이렇게 고안된 하드웨어 구조를 ERNIE(Expansible & Reconfigurable Neuro Informatics Engine)라 명명하였다.

2. SPE(Synapse Processing Element)

SPE는 신경망 연산의 핵심적인 연산을 담당 하고 있으며 곱셈기, 덧셈기, 누산기의 세 가지 블록으로 구성되어 있다. 주요 역할은 퍼셉트론의 서밍 노드 연산과 코호넨 네트워크의 코호넨 노드 연산의 수행이다[9]. 퍼셉트론에서 활성화 함수의 입력을 *sum*, 코호넨 연산에서 구해지는 입력과 가중치와의 유클리드 거리를 *dist*라 하고 **x**와 **w**가 각각 입력과 가중치의 n차원 벡터일 때 이들의 관계는 다음과 같이 표현 할 수 있다.

$$sum = \sum_{i=1}^{n} x_i w_i \qquad (1)$$

$$dist = \sqrt{\sum_{i=1}^{n}(x_i - w_i)^2} \quad (2)$$

그러나 식 (2)의 유클리드 거리를 구하기 위해서는 실수의 제곱 연산 및 제곱근 연산이 필요하며 이러한 연산을 하드웨어로 구현하는 것은 매우 복잡하다. 그리하여 SPE에서는 입력벡터와 가중치 벡터와의 시티-블럭 거리를 계산하여 유사도 측정의 기준으로 삼으며 이는 다음과 같이 표현된다.

$$sum_{SPE} = \sum_{i=1}^{n} x_i w_i + passin \quad (3)$$

$$dist_{SPE} = \sum_{i=1}^{n} x_i - w_i \quad (4)$$

식 (4)에서 $dist_{SPE}$는 오직 정수의 덧셈 연산만을 필요로 하기 때문에 식 (2)와 비교하여 그 연산과정이 매우 간단하다. SPE에 내장되어있는 누산기는 매 클럭마다 컨트롤 신호에 의해 덧셈기의 결과가 양인 경우 값을 더하고 음인 경우 값을 빼는 연산이 가능하여 $x_i - w_i$의 절대값을 별도의 회로 없이 간편하게 계산할 수 있다. 여기서 *passin*은 시냅스 확장의 경우 마스터 관계에 있는 SPE로부터 전달받는 누적합 값을 의미하며 시냅스 확장을 하지 않는 경우나 해당 SPE의 마스터 SPE가 없는 경우에는 *passin* = 0 이 된다. SPE는 내부의 구성 레지스터에 의해 서밍노드 혹은 코호넨 노드로 결정되며 <그림 9>를 통하여 각각의 경우에 있어서 SPE가 동작하는 모습을 확인할 수 있다.

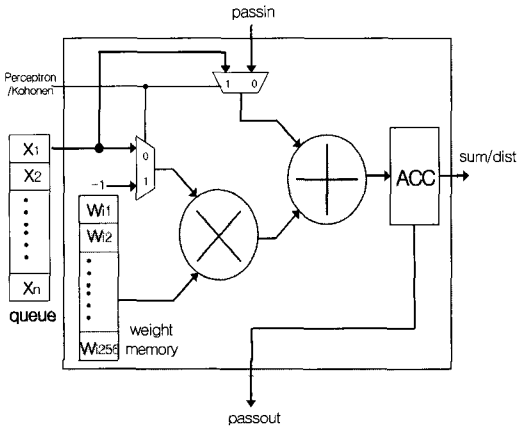


그림 9. SPE의 블록 다이어그램
Fig. 9. Block diagram of SPE.

3. LPE(Layer Processing Element)

LPE는 크게 두 가지 블록으로 나뉜다. 첫 번째 블록은 뉴런의 활성화 함수의 역할을 하는 LUT(Look Up Table)이며, 두 번째 블록은 코호넨 층을 구성하는 뉴런 중에서 승자(winner) 뉴런의 인덱스를 출력하기 위한 KLB(Kohonen Logic Block)이다. 퍼셉트론 연산을 위해 SPE의 출력값인 *sum*을 입력으로 받은 후 이를 LUT의 어드레스로 이용하며 해당 데이터를 활성화 함수의 출력으로 내보낸다. 활성화 함수는 원점을 기준으로 $\pm r$ 의 특정 구간만 고려하며 LUT의 출력값은 sum에 대한 이산함수 F(*sum*)으로 표현된다. LUT의 어드레스가 n bit 이고 x=0 인 순간에만 1의 값을 갖는 단위 임펄스 함수를 $\delta(x)$, 고려하고 있는 활성화 함수를 연속함수 $f(x)$라고 할 때 LUT의 출력 out_{percep}는 다음과 같이 표현된다.

$$sum_{norm} = \left(\frac{sum}{2^{n-1} - 0.5} - 1\right) r \quad (5)$$

$$out_{percep} = F(sum) = \sum_{k=0}^{2^n - 1} f(sum_{norm})\delta(sum - k) \quad (6)$$

코호넨 연산의 경우 KLB는 <그림 10>과 같은 과정을 통하여 승자 뉴런의 인덱스를 출력하게 된다.

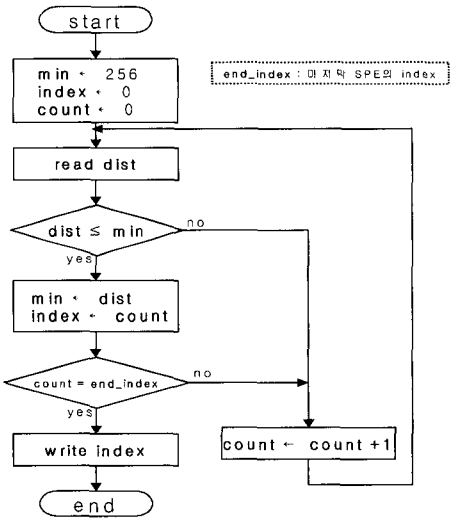


그림 10. KLB에서 승자뉴런의 인덱스를 결정하는 과정
Fig. 10. Process of determining winner neuron in KLB.

<그림 11>은 LPE의 내부 구조에 대한 블록 다이어

그램이며 <그림 12>는 KLB의 블록 다이어그램이다.

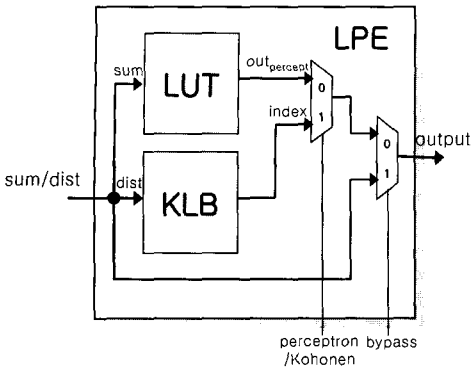


그림 11.LPE의 블록 다이어그램
Fig. 11. Block diagram of LPE.

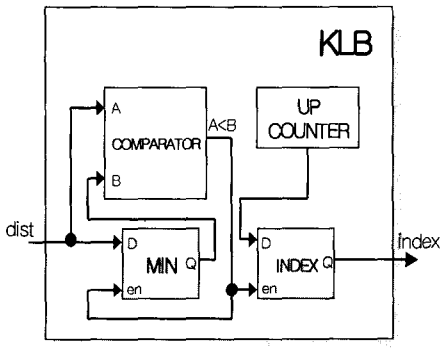


그림 12.KLB의 블록 다이어그램
Fig. 12. Block diagram of KLB.

Ⅳ. ERNIE를 이용한 네트워크 구현

ERNIE에서 특정한 네트워크를 구현하기 위해서는 사용된 MPU를 구성하고 있는 모든 SPE, LPE의 상태와 LUT의 참조 데이터를 결정해 주어야 한다. 따라서 SPE와 LPE의 내부에는 각 유니트의 상태를 규정하는 구성 레지스터가 존재하며 동작 초기에 설정 모드를 통하여 이러한 구성 레지스터 및 LUT의 메모리에 적절한 값을 넣어줌으로써 매우 다양한 신경회로망을 구현하는 것이 가능하다. SPE와 LPE는 각각 4bit, 2bit의 구성 레지스터를 갖고 <그림 13>과 같은 형식을 따른다. <그림 14>는 이러한 구성 레지스터로 인하여 SPE와 LPE가 가질 수 있는 상태를 트리의 형태로 표현한 그림이다.

모든 유니트들은 규정된 상태에 따라서 적절한 역할을 수행하게 되며 그 결과 단층 퍼셉트론을 비롯하여

Configuration bits of SPE

3	2	1	0
Used/ Unused	Perceptron/ SOM	Start	End

Used/Unused : Used SPE or Unused SPE
Perceptron/SOM : Perceptron node or SOM node
Start : Start block of neuron
End : End block of neuron

Configuration bits of LPE

1	0
Mode[1]	Mode[0]

00 : SOM mode
01 : Data passing mode
10 : Perceptron mode
11 : Nothing

그림 13. SPE와 LPE의 구성 레지스터
Fig. 13. Configuration register of SPE and LPE.

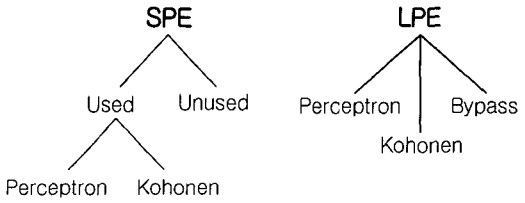


그림 14.SPE와 LPE의 상태 트리
Fig. 14. State tree of SPE and LPE.

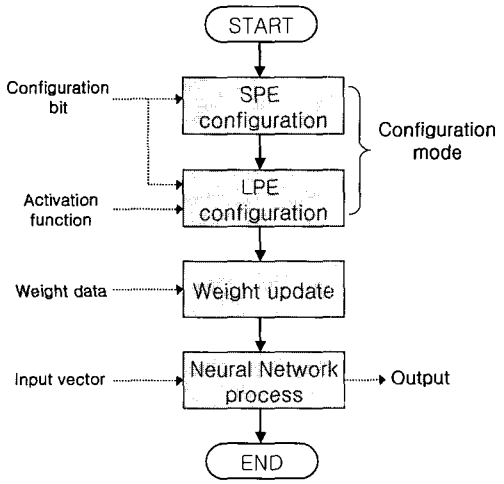


그림 15.ERNIE의 동작 흐름도
Fig. 15. Flowchart of ERNIE operation.

두 개 이상의 은닉층을 가지는 다층 퍼셉트론은 물론 코호넨 층을 구성하여 SOM을 구현할 수도 있다. 더욱

이 퍼셉트론의 경우 비선형 함수의 출력값이 특정한 연산을 통해서 구해지는 것이 아니라 일종의 참조 테이블(Look up table)을 통해 얻어지는 것이기 때문에 사용 가능한 활성화 함수의 제약이 존재하지 않는다.

이와 같이 ERNIE는 확장성 및 범용성을 고려한 설계로 인하여 매우 다양한 종류의 신경회로망을 동일한 플랫폼(platform)위에서 별도의 노력 없이 구현할 수 있으며 상황에 따라서는 퍼셉트론과 SOM이 결합되어 있는 형태의 하이브리드(hybrid) 신경회로망을 구현하는 것 역시 가능하다. <그림 15>는 ERNIE를 이용하여 네크워크를 구성한 뒤 가중치를 저장하고 입력 벡터를 적용하여 출력 값을 확인하기까지의 전체적인 동작 과정을 나타낸 흐름도 이다.

V. 성능분석

하나의 입력패턴에 대하여 출력패턴이 나오기까지의 시간 t(clock cycle)는 식 (7)과 같이 표현된다.

$$t = \sum_{i=0}^{l-1} n_i s_i + (l-1)c + m \tag{7}$$

n_i : i번째 레이어의 뉴런수
l : 레이어의 개수
c : SPE의 연산시간(clock cycle)
m : MPU의 개수

s_i는 네트워크의 시냅스 확장과 관련된 파라미터로서 식 (8)과 같은 범위를 만족하는 정수이다.

$$s_i - 1 < \frac{n_i}{h} \leq s_i \tag{8}$$

여기서 h는 1개의 SPE가 처리할 수 있는 시냅스의 수를 나타낸다. 식 (7)로부터 <그림 16>, <그림 17>과 같은 그래프를 얻을 수 있다.

<그림 16>에서 은닉층의 뉴런과 출력뉴런이 많아질수록 연산시간은 선형적으로 늘어나는 것을 확인할 수 있다. 일반적인 신경회로망은 뉴런의 증가에 따라 연산횟수가 지수적으로 증가한다. 그러나 ERNIE에서는 병렬적으로 연산이 수행되기 때문에 연산횟수의 지수증가에 대하여 연산시간은 선형적으로 증가한다.

<그림 17>은 20MHz로 동작시킬 때 ERNIE의 연산속도를 보여주는 그림이다. 지수적인 연산횟수에 대해

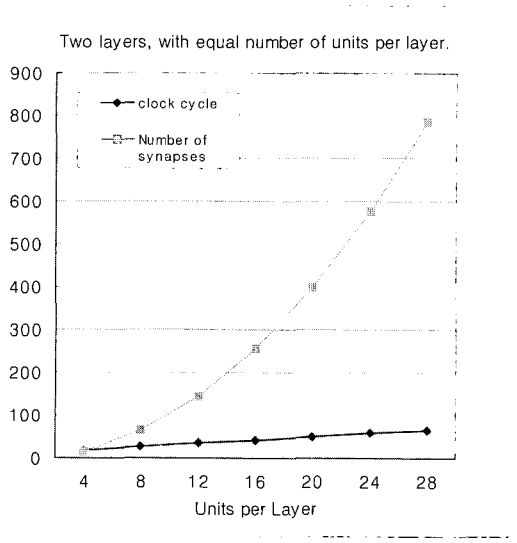


그림 16. 시냅스 증가에 대한 ERNIE의 연산시간
Fig. 16. Execution time caused by adding synapses.

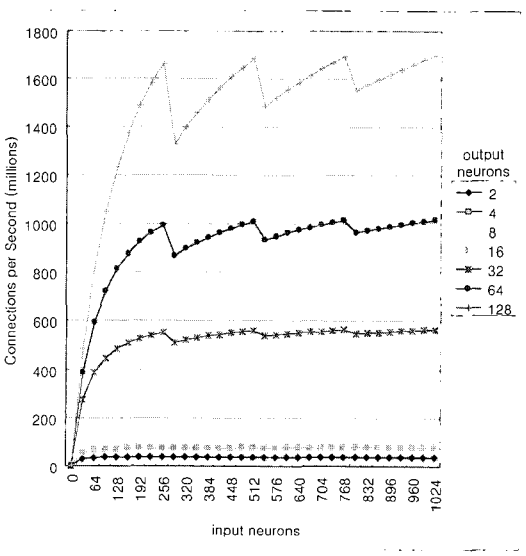


그림 17. ERNIE의 연산속도(CPS)
Fig. 17. Execution speed of ERNIE.

연산시간이 선형적으로 증가한다는 것은 바꿔 말하면 연산횟수가 늘어날수록 연산속도는 빨라진다는 말과 같다. 따라서 뉴런의 수가 증가할수록 연산속도는 점점 증가하게 된다. 출력뉴런의 수가 고정되어 있는 경우 입력뉴런의 수가 256의 배수인 경우에 연산속도는 극대값을 가지며 128개의 출력뉴런을 사용할 경우 1초에 최대 17억 번 정도의 시냅스 연산이 가능하다는 것을 알 수 있다.

<그림 17>에서 보이는 톱니모양은 입력뉴런이 많아지는 데에 따른 시냅스 증가로 인하여 두개 이상의 SPE가 하나의 뉴런을 구성함으로써 그만큼 병렬성이 감소하고 이것이 연산시간의 증가 및 연산속도의 감소로 이어지기 때문에 나타나는 현상이다.

참고로 <표 3>은 Adaptive Solutions에서 개발한 SIMD 구조의 상업용 신경망 칩인 CNAPS와 ERNIE를 구성능력과 최고성능의 측면에서 비교한 것이다[10].

표 3. ERNIE와 CNAPS와의 성능 비교
Table 3. Performance comparison with ERNIE and CNAPS.

구 분	Maximum Configuration	Peak Performance
ERNIE	4 chips 120 PUs/chip	6.3 GCPS
CNAPS	8 chips 64 PUs/chip	5.8 GCPS

VI. 실험방법 및 결과

제안된 ERNIE를 Verilog HDL을 이용하여 설계하였다. 설계된 회로의 동작을 검증하기 위하여 신경망을 구성하는데 필요한 데이터를 HDL 검증과 C언어로 작성된 모델링 검증에 동일하게 적용시켰으며 그 실험결과를 서로 비교하였다. <그림 18>은 이러한 검증 과정을 보여주고 있는 그림이다. HDL 검증은 Mentor Graphics사의 Modelsim 5.5 SE를 이용하여 타이밍 수준에서 수행 되었다.

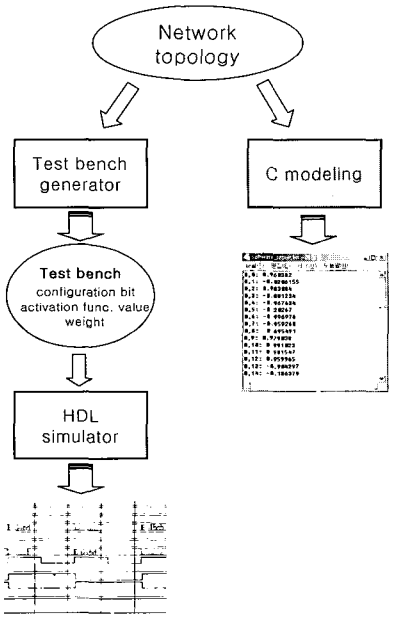


그림 18.ERNIE의 동작 검증 과정
Fig. 18. Evaluation process of ERNIE.

ERNIE의 범용성과 확장성을 증명하기 위하여 다양한 종류의 퍼셉트론 네트워크 및 코호넨 네트워크를 구성하고 연산결과를 확인하였다. 모두 5가지 종류의 신경회로망을 구성하였으며 각각의 실험결과는 시냅스, 뉴런, 레이어의 확장이 가능하다는 것을 보여준다. 퍼셉트론의 경우 ERNIE의 연산결과와 C 모델링 결과와의 평균오차 *error*은 (9)와 같은 수식으로 구하였으며 이상의 내용을 <표 4>와 같이 정리 하였다.

$$error = \frac{\sum_{i=1}^{T}\sum_{j=1}^{N} oute_{ij} - outc_{ij}}{T \times N} \quad (9)$$

T : 테스트 패턴의 개수
N : 출력층 뉴런의 개수
$oute_{ij}$: i번째 테스트 패턴에 대한 j번째 출력 뉴런 연산 결과(ERNIE)
$outc_{ij}$: i번째 테스트 패턴에 대한 j번째 출력 뉴런 연산 결과(C 모델링)

표 4. ERNIE를 이용해 구현한 신경회로망
Table. 4. Neural network implemented using ERNIE.

종 류		MPU	*error*	비 고
퍼셉트론	5 * 20	1	0.0147	MPU를 이용한 신경망 구성
	5 * 40	2	0.0150	뉴런의 확장
	300 * 15	2	0.0291	시냅스의 확장
	256*96*26	7	0.0435	레이어의 확장
코호넨	8 * 8 (코호넨층)	3	-	8*8의 코호넨층 구성

ERNIE를 이용하여 간단한 영문자 알파벳 인식 실험을 수행하기위한 신경회로망을 구성하였다. 신경회로망 구조는 <그림 19>에서 보는 바와 같이 입력층, 은닉층, 출력층으로 구성된 다층 퍼셉트론을 사용하였다. 하나의 알파벳 문자를 픽셀 당 0~255의 값을 가지는 16*16 그레이 스케일의 영상으로 표현한 후 픽셀 값을 -1~1로 정규화 하여 입력으로 사용하며 96개의 뉴런

으로 은닉층을 구성하였다. 또한 문자 인식 결과를 One-hot 인코딩 방식으로 표현하여 출력층은 26개의 뉴런을 갖는다.

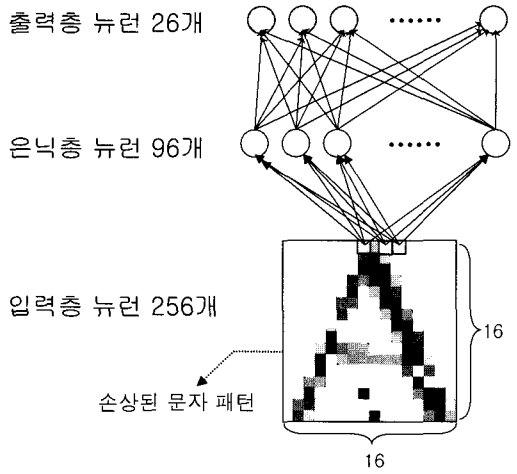


그림 19. 알파벳 인식을 위한 신경회로망 구조
Fig. 19. Neural network structure for alphabet recognition.

제안된 구조에는 학습 모듈이 내장되어 있지 않기 때문에, PC상에서 C언어로 구현된 학습 모델을 통해서 학습된 가중치를 다운로드하여 사용하는 pretrained 방법을 사용하였다. 입력 패턴으로는 영어 대문자 26개에 대해서 5가지 폰트를 사용했으며 정상패턴으로 학습을 시킨 후 손상된 패턴을 입력하는 방법으로 인식 능력을 검증 하였다. <표 5>는 학습과 테스트에 사용된 5가지 폰트의 입력 패턴에 대한 예이다.

표 5. 알파벳 인식에 사용된 학습패턴 및 테스트 패턴의 예
Table. 5. Example of training pattern and testing pattern for alphabet recognition.

구 분	오이체	엽서체	바탕체	그래픽체	굴림체
학습 패턴	A B C	A B C	A B C	A B C	A B C
테스트 패턴	A B C	A B C	A B C	A B C	A B C

각 SPE에 저장되는 가중치는 소프트웨어 학습 알고리즘으로 구하였다. 학습율은 0.05이며, 에러 임계값 보다 에러가 작아지면 학습을 중지 시키는 방법을 사용하였다. 이때 학습에 걸린 반복 횟수는 약 2000회 였다. <표 3>에서 나타낸바와 같이 다층 신경회로망을 구성하기 위하여 MPU 모듈 7개를 사용하였으며 C 모델링 결과와의 평균 오차는 0.0435 로서 매우 근소한 오차를 보였다. 또한 에러 임계 값을 0.001로 주었을 때 HDL 시뮬레이션 결과는 C 모델링의 결과와 동일한 수준의 인식률을 보인다는 것을 확인할 수 있었다.

ERNIE의 코호넨 네트워크 구현 능력을 검증하기 위해 MPU(24 SPE/LPE) 3개를 이용하여 <그림 20>과 같이 8*8의 코호넨 층을 구성한 후 패턴 분류기의 벤치마킹용 데이터 세트로 자주 사용되는 Wisconsin Breast Cancer Database (January 8, 1991)[11]의 패턴 분류를 시도하였다.

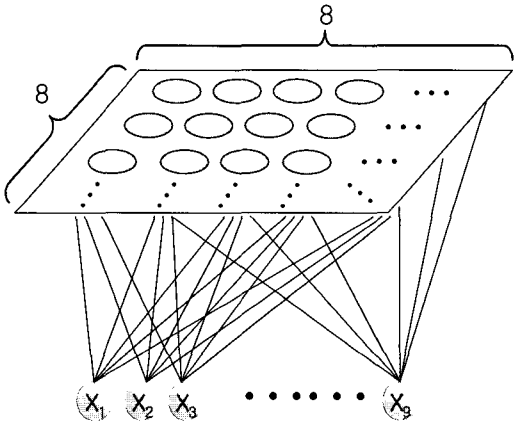
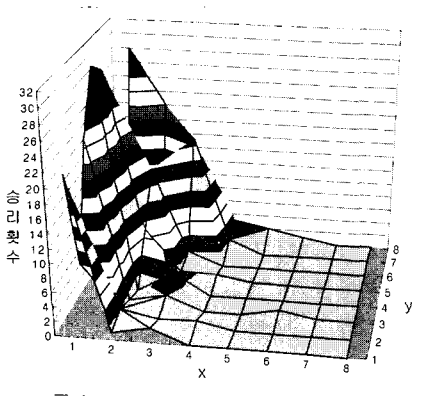


그림 20. 9개의 입력을 받는 2차원 코호넨 층(8*8)
Fig. 20. 2-dimensional Kohonen layer(8*8) with 9 input.

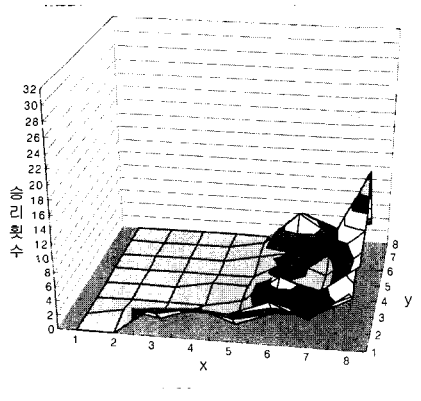
이들 중 정상 패턴 229개와 비정상 패턴 120개를 학습패턴으로 사용하고 나머지 350개의 패턴은 테스트 패턴으로 사용하였다. 학습은 PC에서 수행하였으며 학습율은 학습 횟수 10000회 동안 0.05에서 0.001까지 점진적으로 작은 값을 적용하였다. 학습의 결과로 얻은 가중치를 바탕으로 테스트 패턴을 C 모델링과 ERNIE를 이용한 신경회로망에 동일하게 적용시킨 결과는 <그림 21>과 같았다.

<그림 21>에서 x-y 평면은 각 격자가 하나의 뉴런인 코호넨 층을 의미하며 z축은 입력 패턴에 대해 각 뉴런이 승자가 된 횟수를 나타낸다. 정상 패턴이 입력되었을 때 ERNIE와 C 모델링 모두 코호넨 층의 좌측 영역의 뉴런을 활성화 시켰으며 비정상 패턴이 입력되었을 때에는 우측 영역의 뉴런을 활성 시켰다는 것을

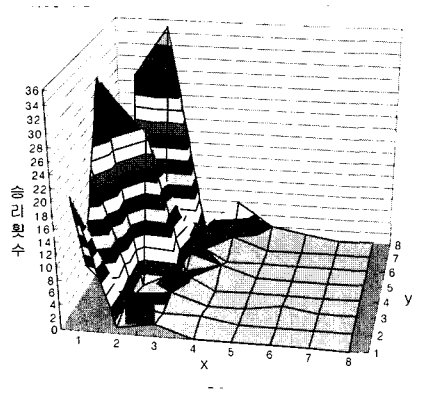
확인할 수 있다. 이 실험을 위한 코호넨 연산에서는 하나의 패턴이 입력될 때 마다 63번의 비교 연산을 통해 승자 뉴런이 결정되는데 350개의 테스트 패턴에 대하여 ERNIE와 C 모델링과의 비교 연산 결과가 다르게 되는 경우는 최대 166번이었다. 따라서 코호넨 층의 뉴런수를 N, 테스트 패턴의 수를 I라고 하면 비교 연산 오차율 err_{comp}는 다음과 같다.



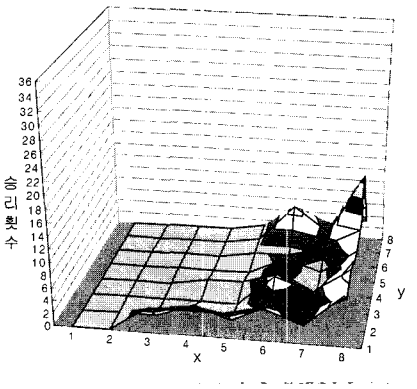
(a)



(b)



(c)



(d)

그림 21. 코호넨 레이어 상에서 각 뉴런의 승리횟수 분포도 ERNIE : (a) 정상 패턴 (b) 비정상 패턴 C 모델링 : (c) 정상 패턴 (d) 비정상 패턴

Fig. 21. Distribution chart of each neuron's winning number in Kohonen layer ERNIE : (a) benign (b) malignant C modeling : (c) benign (d) malignant

$$err_{comp} = \frac{166}{(N-1)I} \times 100$$
$$= \frac{166}{(64-1)\times 350} \times 100$$
$$\approx 0.75\,\% \quad (10)$$

또한 이 경우 서로 다른 비교 연산 결과로 인하여 승자 뉴런이 달라질 경우는 최대 47번이었으며 이 때의 오차율 err_{win}는 다음과 같이 표현된다.

$$err_{win} = \frac{47}{I} \times 100$$
$$= \frac{47}{350} \times 100 \approx 13.43\,\% \quad (11)$$

식 (11)은 C 모델링과 비교하여 그 결과가 최대 13.43%만큼 틀려진다는 것을 의미하지만 이러한 경우 대부분 서로 인접해 있는 뉴런이 승자가 되므로 코호넨 층에서의 비주얼한 결과를 확인하는 데에는 큰 문제가 되지 않는다.

Ⅶ. 결 론

본 논문에서는 신경망 연산기를 디지털 하드웨어로

구현함에 있어서 발생하는 면적 증가 문제 및 네트워크의 확장을 비롯한 재구성 문제를 효과적으로 개선하기 위해서 제안된 ERNIE를 이용하여 여러 가지 실험을 수행 하였고, 이상의 실험에서 C 모델링 검증의 결과와 HDL 검증 결과와의 오차가 신경회로망의 성능에 거의 영향을 미치지 않는 매우 근소한 범위 내에서만 발생한다는 것을 확인 하였다. ERNIE는 활성화 함수의 종류에 대한 제약이 없고 모듈러 구조를 통하여 얼마든지 확장이 가능하며 각 유니트의 상태에 따라 두 가지 종류의 신경회로망 연산이 가능하기 때문에 다양한 종류의 퍼셉트론 네트워크 및 코호넨 네트워크를 구성할 수가 있다. 이러한 재구성 과정은 단지 설정 모드에서 적절한 구성비트를 입력해주는 것만으로 가능하다. 또한 SIMD 구조를 적용하여 하드웨어의 병렬성을 최대로 높였으며 모든 연산을 파이프 라인의 형태로 수행하게 함으로써 연산 속도의 측면에서도 그 성능을 극대화 시켰다. 이와 같은 구조적 특성으로 인하여 ERNIE는 다목적 신경회로망 하드웨어 구조로서 매우 다양한 분야에서의 응용이 가능하다. 또한 ERNIE가 가지는 가장 중요한 특징 중의 하나는 동작 중에 재구성이 가능하다는 것이며 이를 이용하면 실시간 재구성이 필요한 진화 하드웨어 분야에서 그 효용성이 매우 클 것이라 기대된다.

ERNIE에는 퍼셉트론 네트워크 및 코호넨 네트워크의 학습을 위한 기능이 구현되어 있지 않으며, 온 보드(On-board) 학습을 가능케 하기 위해서는 별도의 학습 모듈이 필요하다. 학습 모듈은 ERNIE의 출력을 입력으로 받아서 저장된 목적 값과 비교한 후 적절한 가중치의 변화량을 산출하는 기능을 수행해야 하고 전체적인 학습 루프를 통제하는 컨트롤러의 역할을 해야 할 것이다. 이러한 학습 모듈과 진화를 위한 부가적인 부분이 보강된다면 ERNIE는 수많은 공학적 분야에서 훨씬 더 유용하게 사용될 수 있을 것이다.

참 고 문 헌

[1] Robert J. Schalkoff, "Artificial neural networks", McGraw-Hill, 1997.

[2] Yingang Wang, Zuocheng Ma, Huaxiang Lu, Shoujue Wang, "Discussion on the methodology of neural network hardware design and implementation", Solid-State and Integrated-Circuit Technology, 2001. Proceedings. 6th International Conference on, vol.1 pp. 113-116, 2001.

[3] Miroslav Skrbek, "Fast neural network implementation", Neural Network World, pp. 357-391, 1999.

[4] Tams Szab, Lrinc Antoni, Gbor Horvth, Bla Fehr, "A full-parallel digital implementation for pre-trained NNs", Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on 2000, vol.2, pp. 49-54, 2000.

[5] Masahiro Murakawa, Shuji Yoshizawa, Isamu Kajitani, Xin Yao, Nobuki Kajihara, Masaya Iwata and Tetsuya Higuchi, "The GRD Chip : Genetic Reconfiguration of DSPs for Neural Network Processing", IEEE Transaction on computers, vol.48, no.6, June 1999.

[6] Bernard Girau, "Digital hardware implementation of 2D complatible neural networks", Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on 2000, Volume: 3, pp. 506-511, 2000.

[7] B. Pino, F.J.Pelayo, J. Ortega and A. Prieto, "Design and Evaluation of a Reconfigurable Digital Architecture for Self-Organizing Maps", Microelectronics for Neural, Fuzzy and Bio-Inspired Systems, MicroNeuro '99. Proceedings of the Seventh International Conference on 1999, pp. 395-402, 1999.

[8] S. Vitabile, A. Gentile, G.B Dammone, F. Sorbello, "Multi-layer perceptron mapping on a SIMD architecture", Neural Networks for Signal Processing, Proceedings of the 2002 12th IEEE Workshop, pp. 667-675, 2002.

[9] Simon Haykin, "Neural networks a comprehensive foundation", Prentice hall, pp. 135, pp.448, 1999.

[10] Nikola B. Serbedzija, "Simulating Artificial Neural Networks on Parallel Architectures", Computer, vol. 29, Issue: 3, March 1996, pp. 56-63.

[11] O. L. Mangasarian and W. H. Wolberg : "Cancer diagnosis via linear programming", SIAM News, vol.3, Number 5, September 1990.

저 자 소 개

金 榮 柱(學生會員)

2002년 : 인하대 전자전기컴퓨터공학부 전기 및 제어 전공 졸업. 2004년 2월 : 인하대 정보통신대학원 졸업예정(석사).

童 聖 秀(正會員)

1990년 : 인하대 전기공학과 졸업. 1992년 : 동 대학원 전기공학과 졸업(석사). 1996년~2000년 : 삼성전자 정보통신 네트웍사업부 선임연구원. 2001년~현재 : 용인송담대학 디지털전자정보과 전임강사

李 鍾 浩(正會員)

1976년 : 서울대 전기공학과 졸업. 1978년 : 동 대학원 전기공학과 졸업(석사). 1986년 : 미국 아이오와 주립대 전기 및 컴퓨터 공학과 졸업(공학박사). 1979년~1982년 : 해군사관학교 전임강사. 1980년~1982년 : 국방과학연구소 위촉연구원. 1986년~1989년 : 미국 노틀담대학교 조교수. 1991년~1993년 : 대한전기학회 컴퓨터 및 인공지능연구회 간사장. 1994년~1995년 : 미국 브라운대학교 방문교수. 1997년~1998년 : 인하대 집적회로설계센터소장. 2000년~현재 : 수퍼지능기술연구소 소장. 1989년~현재 : 인하대학교 정보통신공학부 교수.