

論文96-33B-1-16

패턴인식용 VLSI 펄스형 디지털 다계층 신경망의 구조 및 동작 특성

(A VLSI Pulse-mode Digital Multilayer Neural Network for Pattern Classification: Architecture and Computational Behaviors)

金永哲*, 李貴相**

(Young Chul Kim and Gyu Sang Lee)

요 약

대규모 병렬처리가 가능하고 칩당 뉴런 집적도가 높은 펄스형 디지털 다계층 신경망 구조를 제안하였다. 제안된 신경망에서는 대수적인 신경망연산이 의사-랜덤 펄스 시퀀스(pseudo-random pulse sequences)와 단순 디지털 논리 게이트를 이용하여 확률적 프로세스로 대체되었다. 확률적 프로세스의 결과로 나타나는 신경망 연산의 통계적 모델을 제시하였으며 이를 바탕으로 랜덤잡음의 영향과 연산의 정확도를 분석하였다. 이진인식 문제를 적용하여 제안된 신경망의 성능을 평가하고 제시한 통계적 분석결과의 정당성을 검증하였다. Gate 레벨과 register transfer 레벨로 기술된 신경망의 VHDL 모델의 시뮬레이션 결과는 개발된 통계적모델로 예측된 인식추정치와 실제 인식률이 거의 일치함을 보였으며, 또한 숫자인식률에 있어서도 일반 Back-Propagation 신경망의 인식률과 거의 차이가 없음을 보였다.

Abstract

In this paper, a pulse-mode digital multilayer neural network with a massively parallel yet compact and flexible network architecture is presented. Algebraic neural operations are replaced by stochastic processes using pseudo-random pulse sequences and simple logic gates are used as basic computing elements. The distributions of the results from the stochastic processes are approximated using the hypergeometric distribution. A statistical model of the noise (error) is developed to estimate the relative accuracy associated with stochastic computing in terms of mean and variance. Numerical character recognition problems are applied to the network to evaluate the network performance and to justify the validity of analytic results based on the developed statistical model. The network architectures are modeled in VHDL using the mixed descriptions of gate-level and register transfer level (RTL). Experiments show that the statistical model successfully predicts the accuracy of the operations performed in the network and that the character classification rate of the network is competitive to that of ordinary Back-Propagation networks.

* 正會員, 全南大學校 電子工學科

(Dept. of Electronics Engineering, Chonnam Nat'l Univ.)

** 正會員, 全南大學校 電算學科

(Dept. of Computer Science, Chonnam Nat'l

Univ.)

* 이 논문은 1994년도 한국과학재단의 핵심전문연구 과제 연구비에 의하여 연구되었음.

接受日字: 1995年8月17日, 수정완료일: 1995年12月13日

I. 서론

인간두뇌의 탁월한 정보처리능력에 관한 한 현대과학은 두뇌의 막대한 정보에 대한 고도의 동시 병렬처리 능력이 개개 뉴런의 정확성과 복잡성에 기인하는 것이 아니라 단순한 정보처리 능력만을 갖는 많은 뉴런들의 총체적이고 복잡한 연결성에 기인한다고 믿고 있으며 바로 이러한 고도의 상호연결된 신경망을 구현할 전용 하드웨어를 구현하는 것이 인공 신경망(ANN: Artificial Neural Network) 분야의 주요 연구 목표의 하나이다.

현재 많은 ANN 모델들은 디지털 컴퓨터를 이용한 소프트웨어 시뮬레이션(simulation)에 의존하고 있으며 비록 병렬 컴퓨터 상에서 수행된다 하더라도 시뮬레이터의 속도(speed)는 디지털 컴퓨터의 순차적 제어 특성과 communication overhead에 기인하여 전용 VLSI ANN의 그것에 크게 못 미치는 실정이다. 따라서 VLSI 아날로그 또는 하이브리드 ANN 하드웨어가 고정 또는 가변 저항기나 비선형 증폭기의 배열을 이용하여 개발되었으며 아날로그 ANN은 칩(chip)상에 높은 뉴런 집적도의 가능성을 보여 주었다.^[1-4] 하지만, 아날로그 연산의 낮은 정확도, 영구적 아날로그 기억소자의 구현상의 어려움, 그리고 아날로그 소자의 양자잡음·온도·외부 I/O 핀 상의 높은 기생 정전용량(parasitic capacitance) 등과 같은 파라미터 변이로 인하여 현재의 아날로그 VLSI 기술로는 대규모 또는 multichip ANN을 구현하기가 대단히 어려운 실정이다.

한편, 디지털 구현 방식은 이해하기 쉽고 잘 발달된 VLSI 설계 기법과 같은 현재의 디지털 VLSI 기술을 십분 이용할 수 있는 장점을 갖고 있다. 하지만, 기존의 디지털 논리연산장치나 프로세싱 소자(processing elements)를 이용하여 전용 VLSI 신경망을 구현할 경우 차지하는 칩 면적이 대단히 커질 뿐만 아니라 뉴런 상호간 복잡한 연결의 구현상의 어려움이 있다. 따라서 높은 뉴런 집적도를 갖는 고속의 디지털 신경망 하드웨어 개발을 위해서는 고도의 병렬성을 갖는 VLSI 신경망 구조와 여기에 적합한 신호 처리 소자의 개발이 필수적이라 하겠다.^[5]

위에서 언급한 VLSI 신경망 구현상의 한계성을 극복하기 위하여 최근에 확률적 연산기법을 이용한 펄스형 ANN 구조가 저자를 포함한 국내외 연구팀에 의해

서 활발히 연구되고 있다.^[6-10] 이들 구조에서는 신경망에서 수행되는 대수연산을 랜덤 펄스 시퀀스(random pulse sequences)를 이용하여 랜덤 프로세스(random process)로 대신하는 것이 특징이다. 단순한 논리게이트와 간단한 디지털 소자를 이용하여 곱셈과 비선형 변환을 수행한다. 이 방법에서는 모든 신호값 x 는 이진 랜덤 펄스 시퀀스 $x_{(n)}$ 상에서 매 클럭(clock)에 펄스가 발생할 확률로 나타낸다. 하지만, 펄스 시퀀스를 이용한 확률적 연산기법은 일정한 샘플링 기간동안 측정되는 연산 결과를 추정할 수 있는 통계적 모델이 없어 랜덤잡음이 전체 신경망 성능에 미치는 영향을 분석하기가 어려워 성능예측이 가능하면서 신뢰도가 높은 펄스형 VLSI 디지털 신경망 개발의 장애요소가 되고 있다.

본 논문은 확률적 연산기법을 이용한 펄스형 디지털 다계층 신경망의 새로운 구조를 제시하고 확률적 연산의 새로운 통계적 모델을 개발하여 이를 바탕으로 신경망에서의 잡음의 영향과 확률적 연산의 정확도를 분석하고 나아가 이진 패턴인식식 신경망의 인식을 예측 모델을 개발하는 연구이다. 개발된 통계적 모델의 정당성은 VHDL 모델링과 시뮬레이션을 통하여 검증하였다.

본 논문의 구성은 다음과 같다. II장에서는 펄스형 디지털 다계층 신경망을 구성하는 기본소자인 시냅스 소자와 뉴런 몸체 소자에서 수행되는 신경망 연산과 확률대수의 유사성에 대하여 설명한다. III장에서는 펄스형 다계층 신경망의 구조를 제안하고 IV장에서는 제안된 신경망에서 수행되는 확률적 연산의 통계적 모델을 개발하고 랜덤잡음의 영향을 분석한다. V장에서는 숫자 인식문제를 펄스형 다계층 신경망에 적용하여 성능평가를 실행하고 개발된 통계적 모델의 정당성을 검증하며 VI장에서 결론을 맺는다.

II. 확률적 연산기법을 이용한 기본 신경망 소자

펄스형 다계층 신경망은 시냅스(SYN)와 뉴런 몸체(RNB)를 기본 소자로 하여 계층적으로 구성된다. w_{ij} 와 v_i 를 각각 임의 계층의 뉴런 i 와 하위 계층의 뉴런 j 사이의 연결강도와 하위 계층의 뉴런 j 의 출력으로 정의한다. SYN 소자는 그림 1에서 보는 바와 같이 랜덤 펄스 발생기(RPG), 레지스터, 그리고 두개의 AND

게이트와 두개의 wired-OR 선으로 구성된다. RPG는 LFSR (Linear Feedback Shift Register)와 디지털 comparator로 구성된다. w_{ij} 를 저장하는 레지스터의 최상위 비트는 부호를 나타내고 나머지 비트들은 크기를 나타낸다. RPG의 출력으로부터 펄스 시퀀스 $w_{i(n)}$ 가 발생되어 두 개의 AND 게이트의 입력으로 전달되며 w_{ij} 의 부호가 양수일 때는 위쪽의 AND 게이트를 통하여 출력 시퀀스 $m_{i(n)}^+$ 가 나오며 음수일 때는 아래쪽에 있는 AND게이트를 통하여 출력 시퀀스 $m_{i(n)}^-$ 가 발생한다.

만약 두 시퀀스 $w_{i(n)}$ 과 $v_{i(n)}$ 에서의 펄스 발생이 통계적으로 독립적이라면 AND 게이트의 출력 시퀀스에서의 펄스 발생확률은 다음과 같다.

$$m_{ij} = P(m_{i(n)}=1) = P(w_{i(n)}=1 \wedge v_{i(n)}=1) = w_{ij}v_{ij} \quad (1)$$

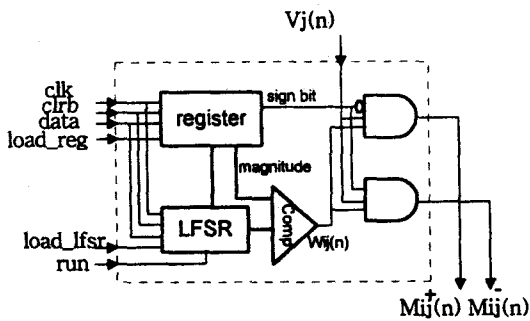


그림 1. 시냅스 소자(SYN)
Fig. 1. Synaptic element(SYN).

위와같이 같은 계층의 시냅스 소자들로부터 발생하는 두 종류의 출력 시퀀스 $m_{i(n)}^+$ 과 $m_{i(n)}^-$ 는 두개의 wired-OR 선인 $net_{i(n)}^+$ 과 $net_{i(n)}^-$ 으로 각각 연결되어 뉴런 몸체 소자 i 로 전달된다.

앞서 설명한 시냅스 소자들로부터 두개의 입력선으로 전달되는 펄스 시퀀스는 AND 게이트를 통하여 RNB내의 계수기(up-counter)에서 수집된다. RNB는 그림 2에서와 같이 AND 게이트, 계수기, 레지스터, 그리고 RPG로 구성되어 있다.

$net_{i(n)}^+$ 와 $net_{i(n)}^-$ 를 각각 뉴런 i 의 흥분성입력과 억제성입력이라고 부른다. $net_{i(n)}^+$ 는 $net_{i(n)}^-$ 과 AND 되어 뉴런 i 의 출력 v_i 를 만든다. 이를 수식적으로 나타내면 다음과 같다.

$$net_i^+ = P(m_i^+=1 \vee m_i^-=1) = 1 - \prod_{j=1}^n (1 - m_{ij}^+) \quad (2)$$

$$= 1 - \prod_{j=1}^n (1 - w_{ij}^+ v_j)$$

같은 방법으로, $net_i^- = 1 - \prod_{j=1}^n (1 + w_{ij}^- v_j)$.

따라서, 만약 $net_{i(n)}^+$ 과 $net_{i(n)}^-$ 이 서로 통계적으로 독립적이라고 가정하면 뉴런 몸체 i 내의 AND 게이트의 출력 시퀀스 $v_{i(n)}$ 에서의 펄스 발생확률은 다음과 같다.

$$v_i = P(net_{i(n)}^+=1 \wedge net_{i(n)}^-=0) = net_i^+ \cdot (1 - net_i^-) \quad (3)$$

$$= [1 - \prod_{j=1}^n (1 - w_{ij}^+ v_j)] \prod_{j=1}^n (1 + w_{ij}^- v_j)$$

식 (3)에서 알 수 있는 바와 같이 v_i 는 연속적이고 미분가능한 비선형 함수이므로 학습을 위하여 역전과 알고리즘을 사용할 수 있다^[11].

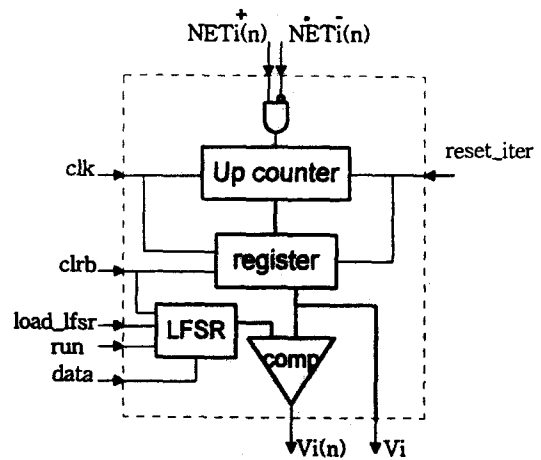


그림 2. 뉴런 몸체 소자(RNB)
Fig. 2. Regular neuron body element(RNB).

그림 2에서 한 주기의 연산이 끝나면 다음 주기의 연산을 위하여 신호 'new_iter'가 '0'에서 '1'로 바뀌며, 이 때 계수기의 출력이 레지스터로 옮겨지고 계수기는 reset된다. 레지스터의 출력은 새로운 출력 시퀀스 $v_{i(n)}$ 을 발생시키기 위하여 RPG에서 사용된다.

III. 모듈형 DMNN 구조

펄스형 디지털 다계층 신경망(DMNN: Digital Multilayer Neural Network)은 위에서 설명한 기본소자를 이용한 다음의 세가지 모듈로 구성된다: 시냅스 배열 모듈(SAM: synaptic array module), 뉴런 몸체 배열 모듈(RNAM: regular neuron body array module), 그리고 상호연결 모듈(ICM: interconnection module). 세가지 모듈을 사용하여 그림 3과 같이 임의 크기의 다계층 신경망을 구성할 수 있는데

전체 신경망 구조는 pipeline 형태로 구성되어 일단 출력 계층에서 출력이 나오기 시작하면 매 연산 반복 주기마다 출력이 나오게된다. 이 신경망 구조의 장점은 같은 계층의 모든 연산은 동시에 이루어지기 때문에 신경망의 크기와 신경망 인식속도는 무관하며 모듈을 통한 네트워크의 확장성이 높은 동시에 칩당 높은 뉴론 집적이 가능하다는 것이다.

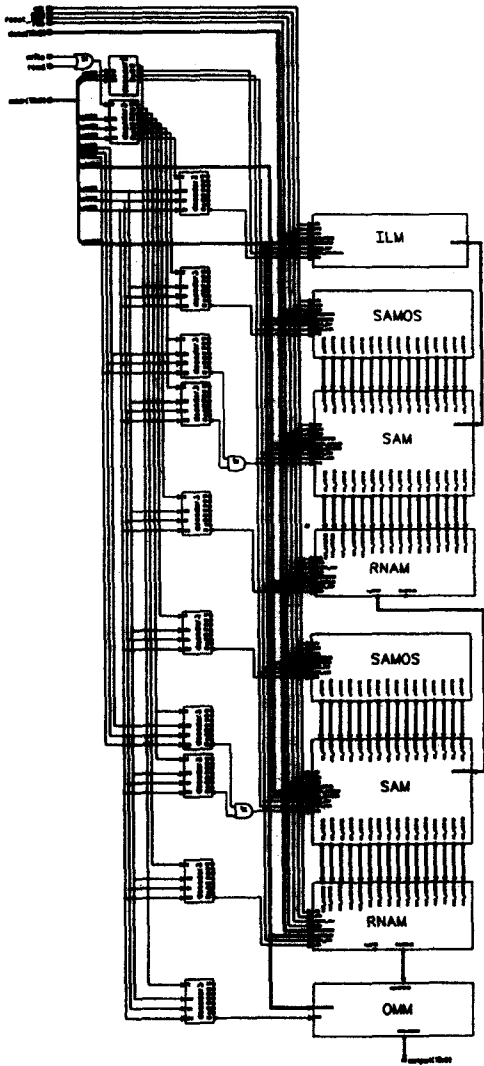


그림 3. 펄스형 디지털 다계층 신경망(8X8X8)의 개략도

Fig. 3. Schematic diagram of a pulse-mode digital multilayer neural network (8X8X8).

SAM은 행렬 형태로 배열된 SYN 소자들과 RNB 입력선들로 구성된다. 연결강도 w_{ij} 는 네트워크 동작이

시작되기 전에 해당 레지스터에 저장되며 각 시냅스 소자내의 $w_{ij(n)}$ 는 하위 계층에서 전달되는 $v_{i(n)}$ 과 함께 AND 게이트의 입력이 되며 같은 계층 내에서는 모든 곱셈은 동시에 수행된다. 그림에서 SAMOS 모듈은 해당계층의 SAM모듈에서 뉴론들의 offset을 위한 SYN 집합을 분리한 모듈이다.

RNAM은 여러개의 RNB로 구성된 모듈이다. SAM에서 전달되는 입력펄스 시퀀스가 AND 게이트를 통하여 계수기에서 카운터되어 $v_{i(n)}$ 을 형성하며 같은 RNAM내의 모든 $v_{i(n)}$ 들은 동시에 발생된다. ICM은 상위 계층으로 뉴론 출력 시퀀스를 전달하는 통로이다. SAM과 RNAM의 자세한 schematics는 논문^[6,7]을 참조할 수 있다.

IV. DMNN 신경망의 통계적 모델과 해석

1. 신호값 x 의 추정치 \hat{x}

펄스형 다계층 신경망에서는 연결강도나 뉴론 출력과 같은 신호값 x 를 그림 1과 2에서 처럼 최대주기(N)를 갖는 LFSR를 포함하는 RPG를 이용하여 펄스 시퀀스 $x_{(n)}$ 상의 '1' 펄스발생 확률로 나타낸다. 이 확률 값은 일정한 클럭 사이클(P_s) 동안 발생한 '1' 펄스(X)를 관측함으로써 추정치 \hat{x} 를 측정한다. $x = X/N$ 이고 $P_s = n$ 이라 할 때 추정치 \hat{x} 의 기대값과 분산은 hypergeometric distribution을 이용하여 각각 다음과 같이 구할 수 있다.

$$E(x) = \frac{1}{n} \sum_{k=0}^N k \cdot P(X=k) = \frac{1}{n} \sum_{k=0}^N k \cdot \frac{\binom{k}{k} \binom{N-k}{N-k}}{\binom{N}{n}} \quad (4)$$

$$= x.$$

$$Var(x) = \frac{1}{n^2} Var(X) = \frac{1}{n^2} (E[X^2] - [E(X)]^2) \quad (5)$$

$$= \frac{x(1-x)}{n} \frac{N-n}{N-1}.$$

2. 첫번째 은닉층에서의 시냅스 곱셈과 비선형 변환 시냅스 소자에서의 AND 게이트의 출력을 대수 표현식으로 나타내면 다음과 같다: $m_{ij} = w_{ij}v_j$. w_{ij} 와 v_j 의 추정치 \hat{w}_{ij} 와 \hat{v}_j 는 첫번째 은닉층에서 원신호와 랜덤 잡음(또는 오차)의 합으로 나타낼 수 있으므로 m_{ij} 의 추정치 \hat{m}_{ij} 는 다음과 같이 쓸 수 있다.

$$\hat{m}_{ij} = (w_{ij} + \Delta w_{ij})(v_j + \Delta v_j) \quad (6)$$

$$= w_{ij}v_j + w_{ij}\Delta v_j + v_j\Delta w_{ij} + \Delta w_{ij}\Delta v_j$$

여기서 Δx_i 는 평균이 제로이고 분산이 $Var(x_i)$ 인 랜덤

잡음이다.

\hat{m}_{ij} 의 기대치와 분산은 각각 다음과 같다.

$$E(\hat{m}_{ij}) = m_{ij} = w_{ij} \cdot v_j, \quad (7)$$

$$Var(\hat{m}_{ij}) = w_{ij}^2 \cdot Var(\hat{v}_j) + v_j^2 \cdot Var(\hat{w}_{ij}) \quad (8)$$

위에서 식 (6)의 이차항은 생략되나 실제로는 아주 적은 상호 상관관계(cross-correlation)를 갖기 때문에 이차항이 완전 제로는 아니다.

펄스형 다계층 신경망이 항상 조건, $P_s = n = N$, 을 만족한다면 신경망내의 연결강도와 뉴런출력을 해당하는 펄스 시퀀스의 평균 펄스 발생율로의 변환 때 생기는 오차를 제거할 수 있다. $n=N$ 일 때 식 (5)에서 $Var(\hat{v}_j) = Var(\hat{w}_{ij}) = 0$ 이므로 식 (8)로부터 $Var(\hat{m}_{ij}) = 0$. 첫번째 은닉층에서 시냅스 곱의 이러한 성질은 확률적 연산 기법의 높은 정확도에 기여한다.

K 번째 은닉층에서의 뉴런 수를 n_k 라 하고 $m_n = m_{i(\max)} = \max [m_{n1}, m_{n2}, \dots, m_{ni}]$ 라 정의하면 n_0 는 입력층의 뉴런수를 나타낸다. n_0 -입력 OR 게이트의 출력시퀀스의 발생 펄스중에는 $m_{n(i)}$ 에서와 동일한 펄스부분이 항상 나타나는 결정론적 성질(deterministic nature)이 존재한다. [6,7]

n -입력 OR 논리 연산의 통계적 모델을 유도하기 위하여 다음의 두 가정을 한다. 이 가정의 타당성은 실험을 통하여 정당화될 것이다. N_{eff}^k 를 k 번째 은닉층에서의 LFSR의 실효주기라 정의하면 $N_{eff}^0 (=N)$ 는 입력층의 LFSR의 실효주기에 해당한다.

가정 1: 첫번째 은닉층에서의 OR 입력수의 증가는 샘플링 주기가 $P_s = N$ 로 고정되고 $m_{ij} = l_{ij}/N = (m_{ij}N_{eff}^1)/N_{eff}^1$ 로 유지될 때 LFSR의 주기 N 을 N_{eff}^1 로 증가시키는 효과와 같다.

가정 2: n_i 가 증가함에 따라 net_i 의 결정론적 비율이 고정값 net_i 와 $0 < \beta \leq 1$ 에 대하여 $m_{i(\max)} = m_n = \max [m_{n1}, m_{n2}, \dots, m_{ni}] = \beta \frac{net_i}{\sqrt{n_0}}$ 와 같이 감소한다.

N_{eff}^1/N 과 β 값은 적용하는 응용문제, 네트워크구조, 그리고 입력패턴에 따라 달라지기 때문에 그 값을 closed form으로 구하기는 불가능하나 이전패턴 인식 문제 응용의 경우 본 연구의 실험 결과에 따르면 $n_0 \geq 2$ 일 때 N_{eff}^1 는 $N + \alpha_1(n_0 - 2)N$ 으로 근사화할 수 있었다. 여기서 α_1 은 첫번째 은닉층에서의 상승계수(incremental parameter)이며 n_0 는 입력층에서의 입력수

이다. α_1 은 0.03과 0.04 사이의 값이며 β 는 1로 가정한다.

n_0 개의 m_{ij} 를 $m_n = m_{i(\max)} = l_n/N$, $m_{i\bar{r}} = 1 - \prod_{j=2}^{n_0} (1 - m_{ij}) = m_{i\bar{r}} N_{eff}^1 / N_{eff}^1$ 의 두 개의 양으로 나누면 랜덤변수 NET_i (시퀀스 $net_{i(n)}$ 에서 샘플링 클럭 주기 P_s 동안 나타나는 '1' 펄스의 수)는 상수 l_n 과 랜덤변수 W 의 합으로 나타낼 수 있다. 즉, $NET_i = l_n + W$ W 는 $N - l_n$ 클럭 사이클 동안에 $m_{i(n)}$, $j \neq 1$,에서의 펄스 발생수를 나타내는 랜덤 변수이다.

n_0 -입력 OR 연산을 위한 W 의 확률함수는 $P(W=k) = \frac{\binom{N-l_n}{k} \binom{N-l_n-k}{N-l_n-k}}{\binom{N-l_n}{N-l_n}}$ 이므로 net_i 의 기대치와 분산은 각각 다음과 같다.

$$E(net_i) = \frac{l_n}{N} + \frac{1}{N} \sum_{k=0}^N k \cdot P(W=k) \quad (9)$$

$$= 1 - \prod_{j=1}^{n_0} (1 - m_{ij}),$$

$$Var(net_i) = \frac{(N-l_n)}{N^2} m_{i\bar{r}} (1 - m_{i\bar{r}}) \frac{N_{eff}^1 - (N-l_n)}{N_{eff}^1 - 1}$$

만약 $m_{i\bar{r}}$ 와 m_n 를 $(net_i - m_n)/(1 - m_n)$ 와 $net_i/\sqrt{n_0}$ 로 각각 치환하면,

$$Var(net_i) = \frac{K_a^1 K_b^1}{N} net_i (1 - net_i), \quad (10)$$

여기서 $K_a^1 = (1 - \frac{1}{\sqrt{n_0}})/(1 - \frac{net_i}{\sqrt{n_0}})$, $K_b^1 = (N_{eff}^1 - N + l_n)/(N_{eff}^1 - 1)$.

3. K 번째 은닉층에서의 시냅스 곱셈과 비선형 변환 K 번째 은닉층에서의 시냅스 곱과 OR 연산을 위한 모델은 앞 절의 유도과정을 일반화한 것이다. K 번째 은닉층에서의 $E(\hat{m}_{ij})$ 와 $Var(\hat{m}_{ij})$ 는 식 (7)과 (8)에서와 같다. 식 (8)의 오른쪽 첫번째 항은 하위층으로부터의 랜덤잡음으로 인하여 제로가 아니므로 $k \geq 2$ 일 때 k 번째 은닉층에서는 $Var(\hat{m}_{ij})$ 는 제로가 아니다. 그리고, 하위계층($k-1$ 번째 은닉층)으로부터의 랜덤잡음의 영향을 $\alpha_k = \alpha_{k-1} + 0.005$ 와 같이 증가한다고 했을 때 N_{eff}^k 는 다음과 같다.

$$N_{eff}^k = N_{eff}^{k-1} + \alpha_k (n_{k-1} - 2) N_{eff}^{k-1} = N_{eff}^{k-1} [1 + \alpha_k (n_{k-1} - 2)]$$

$$= N \prod_{i=1}^k [1 + \alpha_i (n_{i-1} - 2)]$$

여기서 입력층과 출력층은 각각 0 번째 그리고 마지막 은닉층으로 볼 수 있다.

결국, 식 (9)와 (10)의 k 번째의 은닉층에서의 일반형

은 각각 다음과 같다.

$$E(\hat{net}_i) = 1 - \prod_{j=1}^{n_{k-1}} (1 - m_j). \quad (11)$$

$$Var(\hat{net}_i) = \frac{K_a^* K_b^*}{N} net_i (1 - net_i), \quad (12)$$

여기서 $K_a^* = (1 - \frac{1}{\sqrt{n_{k-1}}}) / (1 - \frac{net_i}{\sqrt{n_{k-1}}})$, $K_b^* = (N_{off}^* - N + I_h) / (N_{off}^* - 1)$.

만약 펄스 시퀀스를 통계적으로 독립적인 Bernoulli 시퀀스로 가정한다면 $K_a^* K_b^* = 1$, 그리고 \hat{net}_i 는 이항분포를 갖는다. 식 (10)에서 $i < k$ 에서 n_i 가 대단히 커짐에 따라 $K_a^* K_b^*$ 값은 1로 수렴해가며 $Var(\hat{net}_i)$ 는 이항분포의 분산에 가까워진다.

4. 출력 추정치의 통계적 모델

결국, 임의 계층에서의 뉴런 출력 추정치 \hat{v}_i 의 통계적 모델은 다음과 같다.

$$\begin{aligned} \hat{v}_i &= (net_i^+ + \Delta \hat{net}_i) (1 - net_i^- + \Delta \hat{net}_i) \\ &= net_i^+ (1 - net_i^-) + net_i^+ \Delta \hat{net}_i + (1 - net_i^-) \Delta \hat{net}_i \end{aligned}$$

이차항을 전과 같이 생략하면 k 번째 은닉층에서의 \hat{v}_i 의 기대치와 분산은 각각 다음과 같다.

$$E(\hat{v}_i) = net_i^+ (1 - net_i^-), \quad (13)$$

$$Var(\hat{v}_i) = net_i^{+2} Var(\hat{net}_i) + (1 - net_i^-)^2 Var(\hat{net}_i). \quad (14)$$

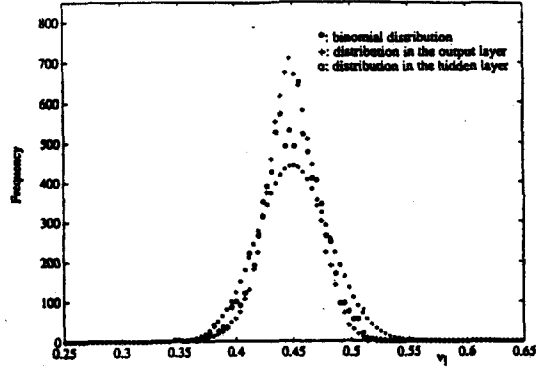
식 (12)에서 n_k 와 k 가 대단히 커짐에 따라 \hat{net}_i^+ 와 $(1 - \hat{net}_i^-)$ 가 각각 이항분포에 가까워지고 $P_i = N$ 이 대단히 클 경우 정규분포에 가까워진다. \hat{v}_i 의 랜덤오차는 식 (14)에서 처럼 \hat{net}_i^+ 와 \hat{net}_i^- 의 랜덤오차의 선형 변환으로 근사화 되므로 \hat{v}_i 또한 정규분포에 가깝다. 그림 4 에서 또는 0.45 (b), $k=2$, $n_0=25$, $n_1=5$, $n_2=5$ $v_i=0.54$ (a) 일 때 시뮬레이션을 통해서 얻어진 출력층 뉴런의 실제 출력분포를 보여준다. 그림으로부터 출력력은 이항분포와 비교해서 편차가 상대적으로 작으면서 정규분포에 가까움을 알 수 있다.

V. DMNN 이진 인식기

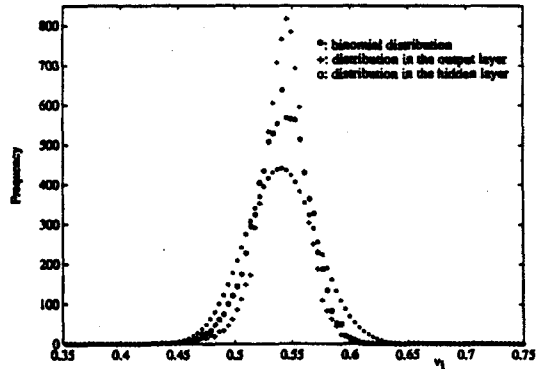
1. DMNN의 추정인식률

앞에서 구한 통계적 모델을 펄스형 다계층 신경망의 해석을 위하여 효과적으로 사용하기 위해서는 입력과 목표 패턴의 표현값뿐만 아니라 계층의 수와 각 계층

의 뉴런 수와 같은 신경망 구성에 관한 정보가 필요하다. 출력 뉴런의 수는 분류할 패턴의 수와 일치하고 0.45 와 0.55 가 출력층 뉴런의 'on'과 'off'를 나타내기 위하여 사용된다. 네트워크는 주어진 입력 패턴에 대하여 오직 한 개의 출력 뉴런만이 'on'이 되도록 학습된다. 학습후 네트워크의 구조는 고정된다.



(a)



(b)

그림 4. (a) $v_i=0.45$, (b) $v_i=0.54$, 그리고 $k=2$, $n_0=25$, $n_1=5$, $n_2=5$ 일 때 8825번의 테스트 결과로 얻어진 은닉층(o)과 출력층(+)에서의 \hat{v}_i 의 분포와 이항분포(*)의 비교

Fig. 4. The distribution of \hat{v}_i in the hidden layer(o) and the output layer(+) for 8825 tests compared to a binomial distribution when (a) $v_i=0.45$, (b) $v_i=0.54$, and $k=2$, $n_0=25$, $n_1=5$, $n_2=5$.

$\hat{v}_i('on')$ 과 $\hat{v}_i('off')$ 가 정규분포를 이룰 때 모든 $j \neq i$ 를 위하여 $\hat{v}_i > 0.5$ 이고 $\hat{v}_i \leq 0.5$ 이면 네트워크는 정확히 입력 패턴을 i 로 분류한다. 다음과 같이 두 랜덤

변수 $V_{on} = \hat{v}_i('on')$ 과 $V_{off} = \hat{v}_i('off')$ 를 정의하면 임의의 입력 패턴이 주어졌을 때 n_k 개의 출력뉴론을 갖는 펄스형 다계층 신경망에서 입력 패턴이 바르게 분류될 확률 P_{cor} 는 다음과 같다.

$$\begin{aligned}
 P_{cor} &= P(V_{on} > 0.5) \prod_{k=1}^{n_k} P(V_{off} \leq 0.5) \\
 &= P(V_{on} > 0.5) P(V_{off} \leq 0.5)^{n_k-1} \\
 &= (1 - F_{V_m}(0.5)) F_{V_{off}}(0.5)^{n_k-1} \\
 &= (1 - F(\frac{0.5-0.55}{\sigma_{V_m}})) F(\frac{0.5-0.45}{\sigma_{V_{off}}})^{n_k-1},
 \end{aligned}
 \tag{15}$$

여기서 $F(u)$ 는 표준 정규 분포함수를 나타내며, 다수의 \hat{v}_i 는 통계적으로 서로 상관관계가 없다고 가정한다.

2. 시뮬레이션

지금까지의 통계적 분석의 정당성을 증명하기 위하여 제안된 펄스형 다계층 신경망의 VHDL 모델을 이용한 시뮬레이션이 수행된다. 두가지의 실험이 실행되었는데 첫째가 0 부터 4 까지 숫자를 분류하는 실험이고 두번째가 0 부터 9 까지 숫자를 분류하는 실험이다. 첫번째 실험에서는 각 숫자의 이미지(image)가 6x4 pixel로 구성되며 두번째 실험에서는 7x5 pixel로 구성된다.

표 1은 세가지 서로 다른 구성의 네트워크에 대하여 식 (15)로부터 얻어진 P_{cor} 를 보여준다. 이 표에서 첫번째 열은 LFSR의 길이(비트수)를 나타내며, 두번째 열은 식 (15)에서 직접 얻을 수 있는 P_{cor} 를, 그리고 세번째 열은 VHDL 시뮬레이션 결과로부터 얻어진 정확한 인식률을 나타낸다. 표에서 첫번째 행은 $n_0 \times n_1 \times n_2$ 의 형태로 네트워크 구성을 나타낸다: 여기서 n_0, n_1, n_2 는 각각 입력층, 은닉층, 그리고 출력층에서의 뉴론수를 나타낸다. 세번째 열에서의 괄호 안의 숫자는 숫자인식이 최대 출력을 갖는 뉴론으로 결정될 때의 정확한 숫자 인식률을 나타낸다. 시뮬레이션을 수행한 결과 2-계층 (한개의 은닉층과 한개의 출력층) 신경망으로써 학습이 가능하였기 때문에 여기서는 2-계층 신경망만을 고려하였다. 세번째 열의 각 숫자는 20,000 번 이상의 테스트를 수행한 결과이며 20번의 다른 초기화 상태로부터의 학습으로부터 얻은 20개의 서로 다른 연결강도 집합을 사용하였으며 각 집합당 1000가지 이상의 서로 다른 LFSR의 초기치 조합을 사용하였다.

표 1 에서 볼 수 있는 바와 같이 펄스형 다계층 신경회로망의 숫자 인식률은 각 네트워크 구성에 따라

P_{cor} 와 비슷한 수치를 보여주며 두 수치상의 차이는 매개변수 α_k 와 β 값이 가장 1과 2에서처럼 일정한 범위 내의 수치로 제한되었다 할 지라도 정확한 값을 구할 수 없기 때문에 생기는 결과이다. 따라서, 이 값들은 N_{off}^* 를 계산하기 위하여 이 범위 내에서 임의로 선택하여야 한다. 표에서의 P_{cor} 를 구하기 위하여 $\alpha_1 = 0.035, \alpha_2 = 0.04, \beta = 1$ 이 사용되었다. 시뮬레이션 결과에서 보는 바와 같이 제시한 통계적 분석 모델이 펄스형 다계층 신경망의 유효한 이진 패턴 인식률을 제공함을 알 수 있다.

표 1. 식 (15) 로 부터 얻은 P_{cor} 와 DMNN 의 VHDL 시뮬레이션으로부터 얻은 인식률

Table 1. P_{cor} obtained from equation (15) and correct classification rates from VHDL simulations.

25 x 5 x 5 (5-숫자 분류)		
	식 (15)로부터 얻은 Pcor	VHDL 시뮬레이션에서 얻은 인식률
8-비트	86.55% ~ 90.91%	92.5%(97.9%)
9-비트	98.59% ~ 99.18%	99.3%(99.65%)

36 x 9 x 10 (10-숫자 분류)		
	식 (15)로부터 얻은 Pcor	VHDL 시뮬레이션에서 얻은 인식률
8-비트	86.64% ~ 92.94%	86.73%(97.8%)
9-비트	97.93% ~ 99.74%	92.6%(99.94%)
10-비트	100%	98.7%(99.06%)

36 x 30 x 10 (10-숫자 분류)		
	식 (15)로부터 얻은 Pcor	VHDL 시뮬레이션에서 얻은 인식률
8-비트	74.25% ~ 83.7%	80.27%(92.4%)
9-비트	96.39% ~ 98.64%	91.2%(99.1%)
10-비트	99.91% ~ 100%	99.1%(99.6%)

VI. 결론

본 논문에서는 확률적 연산기법에 기초한 단일 칩 구현이 가능한 이진 패턴 인식용 펄스형 다계층 신경 회로망 구조를 제안하였으며 네트워크의 통계학적 모델을 개발하여 네트워크 구조에 따른 랜덤잡음이 신경망의 이진 패턴 인식률에 미치는 영향을 분석하였다. 단순한 디지털 소자와 논리게이트로 구성되는 신경망은 각 계층별로 모듈화가 가능하여 모듈을 상호 연결함으로써 임의의 크기로 신경망을 확장할 수 있는 장

점을 갖는다. 나아가, 각 계층의 모든 연산은 병렬처리 되고 각 계층사이의 모든 동작은 파이프라인 방식으로 수행되므로 병렬처리를 극대화시킬 수 있으며 인식 속도가 네트워크의 크기와 무관하고 샘플링 클럭주기와 클럭 속도에만 의존하기 때문에 신경망의 크기에 상관 없이 그 처리속도가 일정하다.

본 논문에서 제안된 펄스형 디지털 다층구조 신경망의 통계학적 모델은 발생확률의 추정치에 따른 랜덤잡음 (또는 오차)의 표준 편차가 Bernoulli 시퀀스를 가질 경우의 그것보다 현저히 적으며 네트워크 크기가 점점 커짐에 따라 이항 분포에 가까워짐을 보였으며 VHDL 모델의 시뮬레이션을 통하여 검증되었다. 이진 패턴 인식 문제를 적용한 실험결과는 해석적 결과에 의해서 제시된 인식률에 대단히 근접하였으며 연산 추정치의 낮은 분산(low variance)의 결과로 목표패턴의 높은 인식률을 보여주었다.

따라서, 펄스형 다계층 신경망이 학습시 낮은 지역 극소점(local minimum)으로 수렴만 이루어진다면 패턴 인식률이 일반 역전파 신경망과 비슷한 고속의 단일 칩 구현이 가능할 것이다. 현재, 본 연구팀은 제안된 신경망의 하드웨어 구현을 위한 VLSI 설계, 신경망 coprocessor의 개발, 패턴의 적절한 코딩기술, 그리고 펄스형 디지털 다계층 신경망의 응용기술 개발 연구를 진행하고 있다.

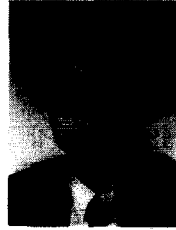
참 고 문 헌

- [1] C. Mead, M. Sivilotti, and M. Emerling, "A Novel Associative Memory Implemented using Collective Computation," *Chapel Hill Conference on VLSI*, pp. 329-342, 1985.
- [2] J.P. Sage and R.S. Withers, "An Artificial Neural Network Intergrated Circuit Based upon MNOS/CCD Principles," *American Inst. of Physics, Neural Networks for Computing*, pp. 381-385, 1986.
- [3] H.P. Graf and L.D. Jackel, "Analog Electronic Neural Network Circuits," *IEEE Circuits and Devices*, pp. 44-49, July 1989.
- [4] M. Holler, S. Tam, H. Castro, and R. Benson, "An Electrically Trainable Artificial Neural Network (ETANN) with 10240 'Floating Gate' Synapses," *Proc. IJCNN*, Vol. II, pp. 191-196, 1989.
- [5] C.E. Atlas and Y. Suzuki, "Digital Systems for Artificial Neural Networks," *IEEE Circuits and Devices*, pp. 20-24, Nov. 1989.
- [6] Y.C. Kim and M.A. Shanblatt, "An Implementable Digital Multilayer Neural Network (DMNN)," *Proc. IJCNN*, Vol. II, pp. 594-600, 1992.
- [7] Y.C. Kim and M.A. Shanblatt, "Random Noise Effects in Pulse-mode Digital Multilayer Neural Networks," *IEEE Trans. on Neural Networks*, Vol. 6, No. 1, pp. 220-229, Jan. 1995.
- [8] A.F. Murray and A.V.W. Smith, "Asynchronous VLSI Neural Networks using Pulse-stream Arithmetic," *IEEE J. of Solid-State Circuits*, Vol. 23, no. 3, pp. 688-697, June 1988.
- [9] W. Wike, D.V. derBout, and T. Miller III, "The VLSI Implementation of STO-NN," *Proc. IJCNN*, Vol. III, pp. 529-534, 1990.
- [10] M.S. Tomlinson Jr, D.J. Walker, and M.A. Sivilotti, "A Digital Neural Network Architecture for VLSI," *Proc. IJCNN*, Vol. II, pp. 545-556, 1990.
- [11] D.E. Rumelhart, G.E. Hilton, and R.J. Williams, "Learning Internal Representations by Error Propagation," *Parallel Distributed Processing*, Cambridge, MA, MIT Press, Vol. 1, pp. 318-362, 1986.

— 저 자 소 개 —

**金永哲(正會員)**

1959년 1월 2일생. 1981년 2월 한양대학교 전자공학과(공학사). 1987년 5월 University of Detroit 전기공학과(공학석사). 1993년 5월 Michigan State University 전기공학과(공학박사). 1993년 8월 ~ 현재 전남대학교 전자공학과 조교수. 주관심분야는 신경회로망 구현 및 응용, VLSI/ASIC 설계, 병렬프로세서 구조, 패턴인식 등임

**李貴相(正會員)**

1958년 2월 1일생. 1980년 2월 서울대학교 전기공학과(공학사). 1982년 2월 서울대학교 컴퓨터공학과(공학석사). 1991년 8월 Pennsylvania State University 전기공학과(공학박사). 1982년 2월 ~ 1983년 3월 금성통신연구소 연구원. 1983년 4월 ~ 현재 전남대학교 전산학과 부교수. 주관심분야는 VLSI 설계자동화, 테스트, 논리합성, 신경회로망 등임