역전파 ANN을 위한 고정 크기 시스톨릭 어레이 설계

김 지 연[0], 장 명 숙[1], 박 기 현[2] (계명대학교)

Design of The Fixed Size Systolic Array for the Back-propagation ANN

Ji-Yeon Kim, Meung-Sook Jang, Kee-Hyun Park (Keimyung University)

Abstract A parallel processing systolic array reduces execution time of the Back-propagation ANN. But, systolic array must be designed whenever the number of neurons in the ANN differ. To use the systolic array which is aready designed as a fixed size VLSI chip, partition of the problem size systolic array must be performed. This paper presents a design method of the fixed size systolic array for thd Back-propagation algorithm using LSGP and LPGS partion method.

Ⅰ. 서 론

인공신경망(Artificial Neural Network, ANN)은 인공지능의 한 분야로 인간의 신경조직을 모델로 하여 단순한 기능을 하는 수많은 노드들을 대규모로 상호연결하는 망 구조를 가진다. ANN 모델은 여러 종류가 있다 이들 중, 다층 퍼셉트론은 입력층과 출력층 사이에 하나 이상의 은닉층을 포함하는 것으로, 역전파 학습 알고리즘(Back- propagation Algorithm)[1]을 이용해서 실생활에 많이 이용되고 있다.

시스톨릭 어레이(Systolic Array)[2]는 동시에 수행되는 규칙적으로 연결된 많은 처리요소(Processing Element, PE)들을 사용하는 병렬컴퓨터 구조의 하나이다. 이는 동기성, 규칙성, 지역성, 선형성 등의 특징을 가지고 있어서, 특히 계산 위주의 문제를 위한 특수 목적의 보조처리기로 사용된다. 역전파 알고리즘은 행렬-벡터 곱셈의 단순한 반복 계산으로, 시스톨릭 어레이로 병렬 처리하면 수행 시간을 줄일 수 있다[3,4]. 그러나 실제로 역전파 알고리즘의 ANN을 시스톨릭 어레이로 설계하기 위하여 수천, 수만개의 PE들이 필요하며, 이것을 VLSI 칩으로 만드는데 어려움이 있다. 또한 입력-은닉-출력층의 노드의 개수가 다를 때마다 서로 다른 시스톨릭 어레이를 설계해야 하는 불편함이 있다[7,8,9]. 그러므로 분할 방법을 사용하면 노드의 수가 다른 역전파 알고리즘을 PE의 개수가 주어진 고정 크기의 시스톨릭 어레이 상에서 수행할 수 있다.

분할(Partion) 방법에는 LSGP(Locally Sequential and Globally Parallel) 방법과 LPGS(Locally Parallel and Globally Sequential) 방법[2,5,6]이 있다. LSGP 방법은 문제의 인덱스 공간을 주어진 고정 크기 어레이의 한 PE에 할당한다. 각 PE에는 대응되는 한 블록 내의 노드들을 순차적으로 수행하게 함으로써, 부분적으로는 순차적인 수행을 하지만 전체적으로는 블록들이 병렬로 수행하게 되는 방법이다. 이와는 대조적으로 LPGS 방법에서는 인덱스 공간이 한 블록을 고정 크기 시스톨릭 어레이 전체에 할당하여 수행시킴으로써 각 블록 내에서는 병렬로 수행이 이루어지지만 블록 전체는 순차적으로 수행하는 방법이다.

0) 계명대학교 컴퓨터 및 전자공학부 석사과정
1) 계명대학교 컴퓨터 및 전자공학부 박사후연구원
2) 계명대학교 컴퓨터 및 전자공학부 교수
본 연구는 97년 정보통신연구관리단 대학기초연구비의 지원을 받았습니다.

본 논문에서는 역전파 ANN을 위해 설계된 병렬처리 시스톨릭 어레이(문제 크기 시스톨릭 어레이)를 LSGP와 LPGS 분할 방법을 사용하여 주어진 고정 크기 시스톨릭 어레이로 설계하는 방법을 제시한다 이를 위하여 주어진 역전파 신경망을 문제 크기의 병렬처리 시스톨릭 어레이로 설계한 후, LSGP 분할 방법을 사용하여, 각 층에 대한 PE 개수와 비례하게 주어진 고정 크기 시스톨릭 어레이의 PE의 개수를 분할하여 각 층에 할당한다. 다음, LPGS 분할 방법을 사용하여, 각 층의 문제 크기 시스톨릭 어레이의 PE들을 여러개의 밴드로 분할하고, 각 밴드를 고정 크기 시스톨릭 어레이에 차례로 사상시킨다. 이렇게 함으로써 수행시간이 증가 되는 단점은 있지만 어떤 크기의 역전파 ANN도 수행 시킬 수 있는 일반적인 고정 크기의 시스톨릭 어레이를 설계할 수 있다.

Ⅱ. 역전파 알고리즘

계층 1에서 계층 L까지의 다 단계 신경망 모델은 계층 1의 입력 계층(input layer), 계층 2에서 계층 $L-1$까지 $L-2$개의 은닉 계층(hidden layer)들, 그리고 계층 L의 출력 계층(output layer)으로 이루어져 있다. 각 층은 인간의 신경세포에 해당하는 뉴런들로 구성되어 있으며, 이웃한 층의 뉴런들은 가중치를 가진 연결선들로 서로 연결되어 있다.

어느 한 입력패턴 I_p에 대한 전방향 연산은, 입력 I_p의 각 원소 $a^1_{i,p}$을 받아들여 각 층에서 활성값을 계산하여 출력을 구하는 단계로, $(l-1)$층의 뉴런의 개수가 L_{l-1}, l 층의 뉴런의 개수가 L_l일 때, 아래의 식 (1)과 (2)로 나타낼 수 있다.

총 입력값 계산

$$u^l_{i,p} = \sum_{j=1}^{L_{l-1}} w^l_{ij}\, a^{l-1}_{j,p} + \theta^l_{i,p} \qquad (1)$$

활성값 계산

$$a^l_{i,p} = f(u^l_{i,p}) = \frac{1}{1 - e^{u^l_{i,p}}} \qquad (2)$$

역방향 연산은, 먼저 출력층의 각 뉴런에 대한 오차 $(t_{i,p} - a^L_{i,p})$ 이 구해진 다음, 이 오차는 아래 식과 같이 출력층에 있는 모든 뉴런 i 에 대하여 제곱하여 더해진다.

총 오차 계산 :

$$E_P = \sum_{j=1}^{M_L} (t_{j,p} - a_{j,p}^L)^2 \quad (3)$$

역전파 알고리즘의 학습 목표는 이 오차값의 평균을 최소화하는 것이다. 이것은 먼저 출력층의 각 뉴런 j 에 대하여 식 (1)을 1차 미분한 뒤 오차에 대한 감소치(error gradient) δ_j^l (l=1, ..., L)을 아래의 식 (4), (4')와 같이 구한다.

오차의 감소치 계산

만약 $l=L$ 일 때

$$\delta_{j,p}^L = (t_{j,p} - a_{j,p}^L)\ f'(u_{j,p}^L) \quad (4)$$

만약 $1 \le l < L$ 일 때

$$\delta_{j,p}^l = \left(\sum_{k=1}^{M_{l+1}} \delta_{k,p}^{l+1}\ w_{kj}^{l+1} \right) f'(u_{j,p}^l) \quad (4')$$

이 오차의 감소치는 역방향으로 신경망의 각 층을 따라 전파되며 식 (5)와 같이 각 층의 가중치를 변경시킨다.

연결선의 가중치 변경 :

$$w_{ij}^l = w_{ij}^l + \eta\ \delta_{j,p}^l\ a_{i,p}^{l-1} \quad (5)$$

여기서 η 는 사용자가 정의한 매개변수(parameter)인 학습 변수(learning rate)이다. 위의 과정은 모든 입력패턴 I 에 대하여 반복되며, 식 (3)의 오차값이 사용자가 정의해 주는 종료 조건인 'ecrit(error criterion)' 보다 작아지면 학습을 멈춘다.

III. 역전파 알고리즘의 문제 크기 시스톨릭 어레이 설계

시스톨릭 어레이를 설계하기 위하여, 전방향, 역방향 식들로부터 자료 흐름을 분석 하여야 한다. 다음, 이 분석을 토대로 자료의 흐름 그래프(Data dependency Graph, DG)를 구한다. 그리고 DG에 공간-시간(space-time transformation) 변환을 적용하여 동시 수행이 가능한 노드들을 시스톨릭 어레이의 서로 다른 PE로 사상시키면 병렬처리가 가능한 문제 크기 시스톨릭 어레이를 설계할 수 있다.[7,8,9] 예를 들어 8-5-2(입력노드 8개, 은닉노드 5개, 출력노드2로 이루어진)역전파 신경망은 그림 1과 같이 1-2 층에 8개, 2-3 층에 2개의 PE를 가지는 문제 크기 시스톨릭 설계할 수 있다.

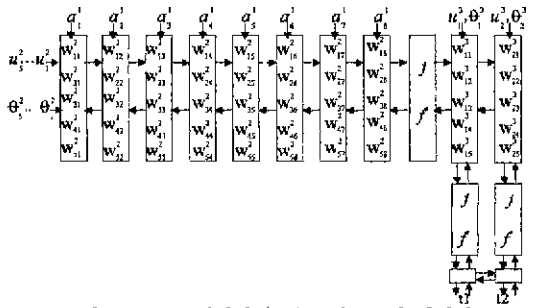


그림 1. 8-5-2 신경망에 대한 시스톨릭 어레이 구조

IV. 역전파 알고리즘의 고정 크기 시스톨릭 어레이 설계

시스톨릭 어레이의 분할 문제는 설계된 문제 크기 시스톨릭 어레이보다 작은 크기의 시스톨릭 어레이(고정크기 시스톨릭 어레이)에 사상시키고자 할 때 제기된다. 즉, 분할이 필요한 이유는 첫째, 문제 크기 시스톨릭 어레이의 크기가 너무 커서 하드웨어로 구현하는데 제약이 있을 경우이고 둘째, 같은 문제를 수행하는데 있어서 특정 크기의 문제를 풀수 있도록 설계된 시스톨릭 어레이로 다른 크기의 문제를 풀수 있게 할 경우이다. 문제 크기 시스톨릭 어레이를 주어진 고정 크기 시스톨릭 어레이로 분할 사상하여 수행하면 분할로 인한 총 수행 시간의 증가를 가져오는 본질적 오버헤드는 감수해야 하지만 분할 방법으로 인한 수행 시간의 증가는 최소화 해야 할 것이다. 즉, 어레이의 분할 문제에 있어서 분할전의 PE의 계산 구조를 그대로 유지하면서 내부 구조가 너무 복잡해지지 않게 하며, 총 실행 시간의 증가를 최소화 하는 사항들을 고려해서 고정 크기 시스톨릭 어레이를 설계해야 한다.

이 절에서는 3절에서 설계된 8-5-2 역전파 신경망에 대한 문제 크기 시스톨릭 어레이를 여러 밴드로 분할한 후, 1-2층에 2개, 2-3층에 1개의 PE를 갖는 고정 크기 시스톨릭 어레이로 사상하여 설계하는 방법을 설명한다. 먼저 LSGP 방법을 이용해서 고정 크기 시스톨릭 어레이의 전체 크기를 문제 크기 시스톨릭 어레이의 각 층의 PE 개수에 비례하게 분할하여 사상시킨다. 고정 크기 시스톨릭 어레이의 각 층에 대한 PE의 개수가 정해지면 각 층에 대해 순서적으로 LPGS 방법을 이용해서 할당된 PE의 개수에 맞게 여러 밴드로 분할한다. 그림 2는 그림 1의 10개의 PE를 가지는 문제 크기 시스톨릭 어레이를 전체 3개의 PE를 가지는 고정 크기 시스톨릭 어레이로 분할하는 방법을 보여준다. 먼저 LSGP 분할 방법을 사용하여, 1-2 층에 8개, 2-3 층에 2개의 PE를 가지는 문제 크기 시스톨릭 어레이를 각 층의 PE 개수에 비례하여, 1-2 층에 2개의 PE를 갖는 밴드 $B0$와, 2-3층에 1개의 PE를 가지는 밴드 $B1$으로 분할 사상한다.

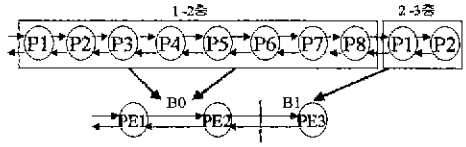
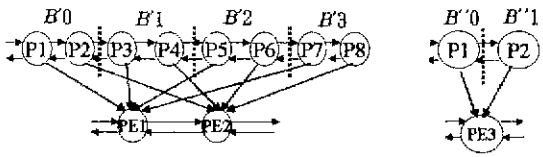


그림 2. LSGP 분할

LSGP 분할 방법으로 각 층의 PE 개수가 정해지면 LPGS 분할 방법을 이용하여 PE가 8개인 1-2 층의 문제 크기 시스톨릭 어레이를 그림 3의 a와 같이 4개의 밴드 $B'0, B'1, B'2, B'3$으로 분할한 뒤 각 밴드를 고정 크기 시스톨릭 어레이의 2개의 PE를 갖는 밴드 $B0$에 사상한다. 같은 방법으로 2-3 층의 2개의 PE를 그림 3의 b와 같이 2개의 밴드 $B''0, B''1$으로 분할한 뒤 각 밴드를 1개의 PE를 갖는 밴드 $B1$의 고정 크기 시스톨릭 어레이로 사상시킨다.



a) 1-2 층 b) 2-3 층

그림 3. LPGS분할

각 층에 대해 LPGS 방법으로 분할된 밴드들이 올바른 수행을 할 수 있도록 각 밴드의 수행 순서를 결정해 주어야 한다. 만약 밴드간의 사이클이 존재한다면 밴드의 수행 순서를 결정하지 못한다. 문제 크기 시스톨릭 어레이의 수행 순서를 보존하기 위해서는 가장 먼저 계산이 일어나는 계산점을 포함하는 밴드를 맨먼저 수행시키고 다음 부터는 밴드간의 의존관계에 의해 수행 순서가 결정된다. 4개의 밴드로 분할된 1-2 층의 시스톨릭 어레이는 분할전 첫

번째 스텝에서 입력값 a_i^1와 가중치 w_{11}^2의 곱셈이 수행되는, 밴드 $B'0$가 맨먼저 수행되고 다음으로 밴드 $B'1, B'2, B'3$ 순으로 계산이 수행되도록 한다. 이 때 밴드 $B'0$의 수행이 끝나면 a_i^2 값은 다음 밴드의 수행에 사용되기 위해 FIFO버퍼에 저장되며, 첫 번째 PE1으로의 귀환 연결선을 갖는 그림 4와 같은 시스톨릭 어레이 구조를 가지게 된다. 입력값 a_i^1 은 L_2번의 반복되는 연산을 수행해야 하므로 회전 레지스터에 저장 되었다가, 역방향 연산시 사용하기 위해 외부 메모리인 LIFO버퍼에 저장된다. 입력 패턴에 대한 활성값을 구하는 비선형 함수 f, f'는 맨 마지막 노드에서 처리해 주도록 설계하여 결과값 a_i^2 은 다시 2-3층의 입력값이 된다. 2개의 밴드 $B'0, B'1$으로 분할된 2-3층에서는, 분할전 첫 번째 스텝에서 수행되는 활성값 a_2^2 와 가중치 w_{11}^3의 계산점을 포함하는 밴드 $B'0$이 먼저 수행되고, 입력값 a_i^2 은 다음 밴드 $B'1$의 수행에 쓰이기 위해 FIFO버퍼에 저장된다. 역방향 연산은 전방향과 대칭으로 역방향으로 진행된다.

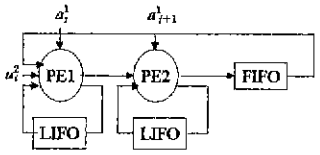


그림 4. 고정 크기 어레이 구조(1-2층)

일반적으로 L_1개의 노드를 갖는 입력층, 각각 $L_2, L_3, \cdots, L_{L-1}$개의 노드를 갖는 은닉층들, 그리고 L_L개의 노드를 갖는 출력층으로 구성된 $L(\geq 2)$층의 역전파 신경망에 대한 분할 크기 시스톨릭 어레이의 각 층(i~$i+1$ 층, 단 $1 \leq i \leq L-1$)이 P_i개의 PE를 가질 때, LPGS 분할에서의 밴드 개수와 고정 크기 시스톨릭 어레이의 총 수행 시간은 다음과 같다.

[정리1] 입력 패턴의 개수를 I_P라 하고 LSGP 분할 방법으로 정해진 고정 크기 시스톨릭 어레이 각 층의 PE 개수를 P'_i라고 할 때

(i~$i+1$) 층에 대한 밴드의 개수는

$$B'_i = \lceil P_i / P'_i \rceil \quad \text{이고,}$$

총 수행 시간은

만약 $L=2$이면 $[I_P \times \{2(B'_1 \times L_2 + (P'_1 - 1) + 2) - 1)\}]$

만약 $L=3$ 이면

$I_P \times [2\{(B'_1 \times L_2) + P'_1 - L_2 + (B'_2 \times L_2) + P'_2 + 1\} - 1]$

만약 $L \geq 4$ 이고 L이 짝수이면

$$I_P \times 2\{(B'_1 \times L_2) + P'_1 + 1 + \sum_{K=2}^{\frac{L}{2}} (B'_{2K-1} \times L_{2K}) + P'_{2K-1} + (B'_{2K-2} \times L_{2K-2}) + P'_{2K-2} - L_{2K-1} + 1\}$$

만약 $L \geq 4$ 이고 L이 홀수이면

$$I_P \times 2\{(B'_1 \times L_2) + P'_1 + 1 + \left\{ \sum_{K=2}^{\lfloor \frac{L}{2} \rfloor} (B'_{2K-1} \times L_{2K}) + P'_{2K-1} + (B'_{2K-2} \times L_{2K-2}) + P'_{2K-2} - L_{2K-2} + 1\} + (B'_{L-1} \times L_{L-1}) + P'_{L-1} - L_{L-1} + 1\}$$

이다.■

앞에서 예를 들은 8-5-2 역전파 신경망의 문제 크기 시스톨릭 어레이를 크기 3인 선형 고정 크기 어레이로 분할 사상하여 설계하면 수행 시간은 57스텝이 된다. 분할전의 수행 시간은 33스텝으로 수행 시간은 약 (밴드수 × 은닉층의 노드수) 즉, (5 × 5=25)만큼 증가한다.

V. 결 론

본 논문에서는 병렬처리에 적합한 역전파 신경망의 문제 크기 시스톨릭 어레이를 LSGP와 LPGS 분할 방법을 이용하여 고정 크기 시스톨릭 어레이를 설계하였다. 먼저 LSGP 분할 방법을 이용하여, 주어진 고정 크기 시스톨릭 어레이의 PE들을 문제 크기 시스톨릭 어레이의 각 층의 PE 개수에 비례하도록 층으로 나눈다. 다음, LPGS 방법을 이용하여, 문제 크기 시스톨릭 어레이의 각 층의 PE들을 여러개의 밴드로 분할한 후 고정 크기 시스톨릭 어레이에 차례로 사상시켜서 역전파 신경망의 고정 크기 시스톨릭 어레이를 설계한다. 각 층에 대해서 LPGS 분할 방법을 이용하므로, 한 밴드 내에서는 노드들이 병렬로 수행되지만 전체적으로는 밴드들이 순차적으로 수행되어 수행 시간의 증가를 가져온다. 그러나 다층으로 구성된 신경망 전체를 보면 LSGP 분할 방법으로 각 층의 PE의 개수를 할당하므로, 분할로 인한 수행 시간의 증가를 어느정도 줄일 수 있다. 또한 LPGS 방법을 사용하여 분할된 여러 밴드들이 주어진 고정 크기 시스톨릭 어레이로 사상되어, 분할전의 PE의 계산 구조를 그대로 유지할 수 있고 내부 구조도 복잡해지지 않았다. 그러나 분할전 보다는 더 많은 외부 메모리가 필요하고 총 수행 시간도 증가하였으나 실제 응용문제에 구현된 고정 크기 시스톨릭 어레이를 크기가 다른 같은 문제에 적용할 수 있는 큰 이점이 있다..

참 고 문 헌

1. Rumelhart, D. E., and McClelland, J. L., *Explorations in Parallel Distributed Processing*, MIT Press. Cambridge, MA.1988.

2. S. Y. Kung, *VLSI Array Processors*, Prentice-Hall, 1987.

3. S. Y. Kung, and J. N. Hwang, "A Unified Systolic Architecture for Artificial Neural Networks," *J. of Parallel and Distrib. Comput.*, No. 6, 1989.

4. S. Y. Kung, and J. N. Hwang, "Parallel Architectures for Artificial Neural Nets," *International Conference on Neural Networks*, San Diego, CA., Vol 2, pp. 165-172, 1988.

5. 이성우, '고정-크기 시스톨릭의 설계 방법," 경북대학교 학위논문, 1996.

6. 이성우, 김윤호, 유기영, "고정크기 시스톨릭 어레이의 설계 방법," 병렬처리시스템 학술발표회 논문집, 제6권, 제2호, pp.31-40. 1995.

7. 장명숙, "역전파 알고리즘의 효율적인 시스톨릭 어레이 설계," 경북대학교 박사학위논문, 1996

8. 장명숙, 박가현, "역전파 ANN을 위한 병렬처리 시스톨릭 어레이," 한국정보 과학회 학술발표 논문집 Vol. 25, No. 1. pp. 623-625, 1998

9. 장명숙, 유기영, "역전파 신경망을 위한 시스톨릭 배열의 설계," 한국 정보 과학회 논문지, Vol.. 24, No. 4, pp.201-212, 1995.