論文98-35C-2-7

다중계층 퍼셉트론 내 Sigmoid 활성함수의 구간 선형 근사와 양자화 근사와의 비교

(A Piecewise Affine Approximation of Sigmoid Activation Functions in Multi-layered Perceptrons and a Comparison with a Quantization Scheme)

尹 炳 文 *, 辛 堯 安 **

(Byung-Moon Yoon and Yo-an Shin)

요 약

비선형 신경회로망 모델인 다중계층 퍼셉트론 (multi-layered perceptron)은 비선형 함수에 대한 우수한 함수 근사 능력 때문에 여러 응용 분야에서 널리 사용되고 있다. 그러나 이 모델을 실제 디지털 시스템으로 구현 시 중간 은닉층 내의 비선형 sigmoid 활성함수를 look-up table (LUT)을 사용한 양자화 방법에 의해 흔히 근사하며, 따라서 정확한 근사를 위해 LUT의 크기가 매우 커야 하는 단점이 있다. 본 논문에서는 이러한 양자화 방법의 단점을 제거하는 방법으로 비선형 sigmoid 활성함수를 몇 개의 적은 개수의 입력 구간으로 나누어 선형 근사하는 "구간 선형 근사 (piecewise affine approximation)" 방법을 제안하고, 이를 이용한 다중계층 퍼셉트론의 표현식과 오차 역전파 학습 알고리즘을 유도하며, XOR 문제에 대한 Monte Carlo 시뮬레이션을 통해 성능을 비교하였다. 실험 결과, 적은 수의 구간을 갖는 구간 선형 근사 방법이 학습 수렴의 측면에서 매우 큰 크기의 양자화 방법보다 월등히 우수함을 알 수 있었으며, 이러한 결과로부터 본 제안 방법이 실제 디지털 하드웨어 구현 시 양자화 방법에 비교하여 저장에 필요한 데이터 수, 근사 오차 그리고 수행 시간의 획기적인 감소가 기대된다.

Abstract

Multi-layered perceptrons that are a nonlinear neural network model, have been widely used for various applications mainly thanks to good function approximation capability for nonlinear functions. However, for digital hardware implementation of the multi-layered perceptrons, the quantization scheme using "look-up tables (LUTs)" is commonly employed to handle nonlinear sigmoid activation functions in the networks, and thus requires large amount of storage to prevent unacceptable quantization errors. This paper is concerned with a new effective methodology for digital hardware implementation of multi-layered perceptrons, and proposes a "piecewise affine approximation" method in which input domain is divided into (small number of) sub-intervals and nonlinear sigmoid function is linearly approximated within each sub-interval. Using the proposed method, we develop an expression and an error backpropagation type learning algorithm for a multi-layered perceptron, and compare the performance with the quantization method through Monte Carlo simulations on XOR problems. Simulation results show that, in terms of learning convergence, the proposed method with a small number of sub-intervals significantly outperforms the quantization method with a very large storage requirement. We expect from these results that the proposed method can be utilized in digital system implementation to significantly reduce the storage requirement, quantization error, and learning time of the quantization method.

* 正會員, LG情報通信(株)
(LG Information & Communications, Ltd.)
** 正會員, 崇實大學敎 電子工學科
(Dept. of Electronic Eng., Soongsil University)

接受日字:1996年12月4日, 수정완료일:1998年2月3日

Ⅰ. 서 론

최근 들어 널리 사용되고 있는 신경회로망 모델인 다중계층 퍼셉트론 (multi-layered perceptron)은 Adaline[1] 등과 같은 선형 모델의 단점인 제한적인 함수 근사 능력을 극복하는 비선형 모델이다[2]. 이러한 비선형적인 특성은 여러 층의 비선형 중간 "은닉층"을 둠으로서 가능하며, 하나의 중간층을 갖고, 중간층 뉴론은 비선형 sigmoid 활성함수 (activation function)[2]를 그리고 출력층 뉴론은 선형 활성함수를 사용하는 다중계층 퍼셉트론은 만약 중간층 뉴론의 개수가 충분히 많다면 우리가 관심 있는 대부분의 비선형 함수를 원하는 오차 내로 근사할 수 있음이 수학적으로 증명되었다[3]~[5]. 다중계층 퍼셉트론을 위한 학습 알고리즘으로서 가장 널리 사용되는 것은 오차 역전파 알고리즘 (error backpropagation algorithm)이다[2]. 이 알고리즘은 지도 학습 (supervised learning) 방법으로서, 선형 모델을 위한 LMS (least mean square) 알고리즘을 다중 계층에 대해 확장한 gradient descent 방법이다. 이 방법은 주어진 목표 출력과 신경회로망의 실제 출력과의 오차를 다중 계층에 역전파 하여 신경회로망 내의 연결 강도들을 적응적으로 개선한다.

다중계층 퍼셉트론을 실제 문제에 적용 시 일반적인 컴퓨터를 사용하여 소프트웨어적으로 구현하거나, 하드웨어 (아날로그, 디지털 혹은 복합적인 방법)[6]로 구현하게 된다. 특히 최근 들어 digital signal processor (DSP)의 획기적인 발전으로 신경회로망을 범용 DSP나 전용 디지털 칩으로 구현하려는 많은 시도가 있어 왔다. 하지만 디지털 하드웨어로 구현 시 이 모델 내에서 사용되는 sigmoid 활성함수의 비선형성을 구현하는데 어려움이 있으며, 이를 해결하기 위해 여러 많은 입력에 대해 이에 해당하는 sigmoid 활성함수의 출력값을 "look-up table (LUT)"의 형태로 저장하여 임의의 입력에 대해 가장 가까운 출력을 이용하는 양자화 방법이 널리 사용된다[7]. 결국 이 방법은 기본적으로 양자화에 따른 오차가 수반되며, 적절한 수준 이하의 양자화 오차를 유지하기 위해서 LUT의 크기가 커져야 하는 단점이 있다. 또한 오차 역전파 알고리즘의 개선식에서는 비선형 sigmoid 함수의 미분값을 이용하게 되며, 이 미분 역시 비선형 함수가 되므로 각 입력마다 해당되는 sigmoid 함수의 미분값을 역시 저장하여 구현의 복잡도와 수행 시간을 상당히 증가시키는 요인으로 작용한다. 특히 학습 알고리즘 개선식은 일반적인 학습 과정에서 매우 많은 반복 동안에 적용되므로, 이러한 복잡도와 수행 시간의 증가는 전체 학습 과정을 상당히 지연시키게 된다.

본 논문에서는 다중계층 퍼셉트론을 디지털 하드웨어로 구현하는 경우, 이 신경회로망 모델 내의 비선형 sigmoid 활성함수를 구현하는 방법으로 널리 사용되는 양자화 방법의 단점, 즉 적절한 수준의 양자화 오차를 위해 많은 량의 저장이 필요하다는 점을 제거할 수 있는 새로운 방법으로서 비선형 sigmoid 활성함수의 입력 범위를 (적은 수의) 여러 구간으로 나누고 각 구간별로 선형 근사하는 "구간 선형 근사 (piecewise affine approximation)" 방법을 제안하고, 이를 이용한 다중계층 퍼셉트론의 표현식과 오차 역전파 학습 알고리즘을 유도하며, XOR 문제에 대한 Monte Carlo 시뮬레이션을 통해 양자화 방법과의 성능을 비교 하고자 한다.

본 논문에서 제안하는 구간 선형 근사 방법은 양자화 방법에서처럼 근사에 따른 어느 정도의 오차를 수반하게 되나, 적절한 구간의 배분으로 이러한 오차를 상당히 줄일 수 있으며, 적은 수의 입력-출력 값들만을 저장하여 디지털 하드웨어 구현 시 탐색 시간을 줄여 실제 수행 시간의 상당한 단축을 얻을 수 있다는 장점이 있을 수 있다.

본 논문의 구성은 다음과 같다. 2 장에서는 비선형 sigmoid 활성함수에 대한 구간 선형 근사 방법을 제안하고, 이를 이용한 다중계층 퍼셉트론의 표현식과 오차 역전파 알고리즘을 유도한다. 이어 3 장에서는 XOR 문제에 대한 Monte Carlo 시뮬레이션 결과를 통해 본 논문의 제안 방법과 양자화 방법의 성능을 비교·분석하며, 이들 방법에 요구되는 저장량, 탐색 시간, 계산량 등을 정량적으로 비교한다. 마지막으로 4장에서 결론을 맺는다.

Ⅱ. Sigmoid 활성함수의 구간 선형 근사 방법

1. 다중계층 퍼셉트론

일반성을 크게 잃지 않으면서 우리가 본 논문에서 고려하는 하나의 중간 은닉층을 갖는 다중계층 퍼셉트론은 아래 그림 1과 같다.

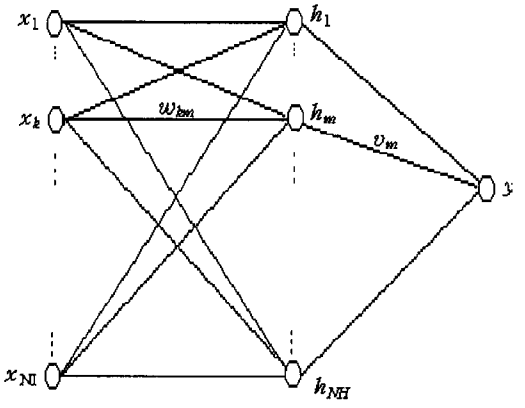


그림 1. 하나의 은닉층을 갖는 다중계층 퍼셉트론
Fig. 1. A multi-layered perceptron with one hidden layer.

여기서 입력은 N_I개, 중간 은닉층 뉴론은 N_H개, 출력층 뉴론은 1개이다. N_I개의 입력은 $x_1, \cdots, x_{N_I}$로 표기되며, $x_0 \equiv 1$은 계산의 편의를 위해 첨가되었다. 유사하게 중간 은닉층 뉴론의 출력은 $h_1, \cdots, h_{N_H}$로 표기되며, $h_0 \equiv 1$ 역시 추후의 계산 편의를 위해 첨가되었다. 학습에 의해 개선되는 뉴론 간 연결 강도(weight connection)는 입력층-중간층의 경우 $(N_I+1)\times N_H$ 행렬로 표현할 수 있으며, 이 행렬의 원소 w_{km} ($k=0,\cdots,N_I$, $m=1,\cdots,N_H$)는 k번째 입력과 m번째 중간층 뉴론 사이의 연결 강도이다. 또한 중간층-출력층의 경우 연결 강도를 $(N_H+1)\times 1$ 벡터로 표현할 수 있으며, 각 원소 v_m ($m=0,1,\cdots,N_H$)은 m번째 중간층 뉴론의 출력과 출력층 뉴론 사이의 연결 강도이다. 중간층 뉴론을 위한 활성함수(activation function)로는 일반적으로 sigmoid 함수를, 그리고 출력층 뉴론을 위한 활성함수로는 문제에 따라 sigmoid 함수나 선형 함수를 사용한다. Sigmoid 함수 $\sigma(\cdot)$는 단조 증가하는 연속 함수로서 아래의 조건을 만족한다[3].

$$\lim_{u\to-\infty}\sigma(u)=a,\quad \lim_{u\to+\infty}\sigma(u)=b\quad (a<b) \tag{1}$$

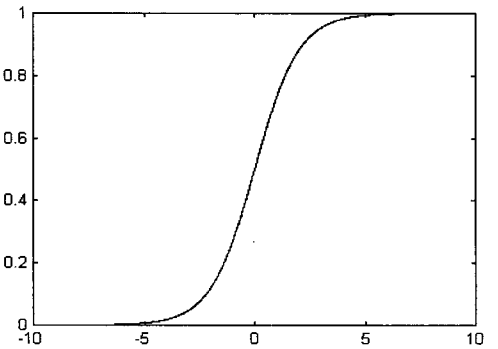
흔히 활성함수 출력의 범위에 따라 다음과 같은 logistic 함수[2]

$$\sigma(u)\equiv\frac{1}{1+e^{-u}}\in(0,1) \tag{2}$$

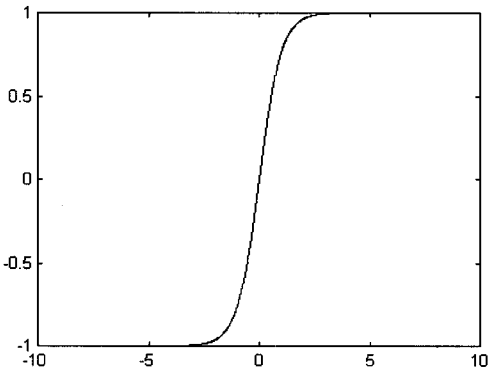
또는 아래의 hyperbolic tangent 함수 $\tanh(\cdot)$를 사용한다.

$$\sigma(u)\equiv\tanh(u)=\frac{e^{+u}-e^{-u}}{e^{+u}+e^{-u}}\in(-1,1) \tag{3}$$

이러한 sigmoid 함수는 일반적으로 S-형태의 입력-출력 포화 특성을 갖게 되며, 다음 그림 2에서 (a)는 logistic 함수를 그리고 (b)는 hyperbolic tangent 함수를 도시한다. 본 논문에서는 식 (3)과 같이 주어지는 hyperbolic tangent 함수를 중간층 뉴론 활성함수로 사용하고, 출력층 뉴론은 $\sigma(u)\equiv u$와 같은 선형 함수를 활성함수로 사용한다고 가정하기로 한다.



(a) Logistic



(b) Hyperbolic tangent

그림 2. Sigmoid 활성함수
Fig. 2. Sigmoid activation functions.

그림 1의 다중계층 퍼셉트론에서 m번째 중간층 뉴론의 입력 net_m은 다음 식 (4)와 같이 표현되고, 이 뉴론의 출력 h_m은 식 (5)와 같이 표현된다.

$$\text{net}_m=\sum_{k=0}^{N_I}w_{km}x_k\ (m=1,\cdots,N_H) \tag{4}$$

$$h_m = \sigma(\text{net}_m) = \sigma(\sum_{k=0}^{N_I} w_{km} x_k) \quad (m = 1, \cdots, N_H) \tag{5}$$

이 때 다중계층 퍼셉트론의 출력 y는 다음과 같다.

$$y = \sum_{m=0}^{N_H} v_m h_m = v_0 + \sum_{m=1}^{N_H} v_m \sigma(\sum_{k=0}^{N_I} w_{km} x_k) \tag{6}$$

2. Sigmoid 활성함수의 구간 선형 근사를 이용한 다중계층 퍼셉트론의 표현식

그림 2(b)와 같은 sigmoid 활성함수는 $u \to \pm\infty$일 때 포화되는 특성을 보인다. 그림 3은 이런 성질을 이용하여 이 활성함수를 구간 선형 근사한 (piecewise affine approximation) 결과를 나타내고 있다. 그림 3에서 임의의 N개 입력 $u_0, u_1, \cdots, u_{N-1}$ ($u_0 < u_1 < \cdots < u_{N-1}$)에 대해 구간 선형 근사된 함수 $\sigma_a(\cdot)$는 아래와 같다.

$$\sigma_a(u) = \begin{cases} -1 & \text{if } u < u_0 \\ +1 & \text{if } u \geq u_{N-1} \\ s_i \cdot [u - u_{i-1}] + \sigma(u_{i-1}) & \text{if } u_{i-1} \leq u < u_i \quad (i = 1, \cdots, N-1) \end{cases} \tag{7}$$

여기서 의미 있는 근사를 위해 다음과 같은 조건을 만족한다고 가정하며,

$$u_0 < 0, \quad u_{N-1} > 0 \tag{8}$$

각 구간 $[u_{i-1}, u_i)$의 선형 근사식의 기울기를 다음과 같이 정의하였다.

$$s_i \equiv \frac{\sigma(u_i) - \sigma(u_{i-1})}{u_i - u_{i-1}} \quad (i = 1, \cdots, N-1) \tag{9}$$

실제 $\sigma_a(\cdot)$는 $u = u_0, \cdots, u_{N-1}$에서 미분값이 존재하지 않으므로 다음과 같은 오른쪽으로부터의 일방향 (one-sided from the right) 미분을 이용하기로 한다.

$$\sigma_a'(u) \equiv \lim_{x \to u^+} \frac{\sigma_a(x) - \sigma_a(u)}{x - u} = \begin{cases} 0 & \text{if } u < u_0 \\ 0 & \text{if } u \geq u_{N-1} \\ s_i & \text{if } u_{i-1} \leq u < u_i (i = 1, \cdots, N-1) \end{cases} \tag{10}$$

식 (7)과 같이 구간 선형 근사된 활성함수 $\sigma_a(\cdot)$를 이용할 경우 출력 y_a는 다음과 같이 구해진다.

$$y_a = \sum_{m:\, u_{i_m-1} \leq \text{net}_m < u_{i_m}} v_m \sigma_a(\text{net}_m) + v_0 + \sum_{m:\, \text{net}_m \geq u_{N-1}} v_m - \sum_{m:\, \text{net}_m < u_0} v_m \tag{11}$$

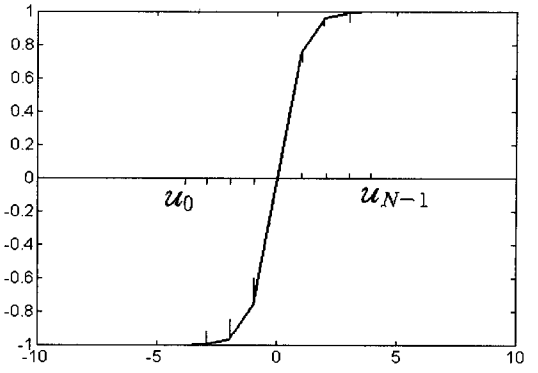


그림 3. Sigmoid 활성함수의 구간 선형 근사
Fig. 3. A piecewise affine approximation of a sigmoid function.

위 식에서 net_m이 갖는 값에 따라 m을 다음과 같이 3 개의 집합으로 구분한다.

$$M^- \equiv \{m \mid \text{net}_m < u_0\} \tag{12-1}$$

$$M^+ \equiv \{m = 0 \text{ or } m \mid \text{net}_m \geq u_{N-1}\} \tag{12-2}$$

$$M^0 \equiv \{m \mid u_{i_m-1} \leq \text{net}_m < u_{i_m}\} \quad \text{where } i_m \in \{1, 2, \cdots, N-1\} \tag{12-3}$$

#$(\cdot)$를 집합에 대한 cardinal number라 하면 $\#(M_0) + \#(M_1) + \#(M_2) = N_H + 1$이다. 결국 식 (7)과 (12)를 이용하면 식 (11)은 아래와 같이 정리된다.

$$\begin{aligned} y_a &= \sum_{m \in M^0} v_m \sigma_a(\text{net}_m) + \sum_{m \in M^+} v_m - \sum_{m \in M^-} v_m \\ &\equiv \sum_{k=0}^{N_I} \left(\sum_{m=0}^{N_H} v_m \alpha_{km} \right) x_k + \sum_{m=0}^{N_H} v_m \beta_m \end{aligned} \tag{13}$$

여기서,

$$\alpha_{km} \equiv \begin{cases} 0 & \text{if } m \in M^- \\ 0 & \text{if } m \in M^+ \\ s_{i_m} w_{km} & \text{if } m \in M^0 \end{cases} \tag{14-1}$$

$$\beta_m \equiv \begin{cases} -1 & \text{if } m \in M^- \\ +1 & \text{if } m \in M^+ \\ \sigma(u_{i_m-1}) - s_{i_m} u_{i_m-1} & \text{if } m \in M^0 \end{cases} \tag{14-2}$$

3. 구간 선형 근사된 sigmoid 활성함수를 이용한 오차 역전파 알고리즘의 유도

이 절에서는 구간 선형 근사된 활성함수를 사용하는 다중계층 퍼셉트론을 위한 오차 역전파 알고리즘을 유

도하기로 한다. n을 반복 시간이라 하면, 이 때의 출력은 식 (13)으로부터

$$y_a(n) = \sum_{k=0}^{N_H} v_m(n)\left(\sum_{m=0}^{N_I} a_{km}(n)\, x_k(n) + \beta_m(n)\right) \tag{15}$$

가 되며, $d(n)$을 이 때의 목표 출력 (desired output)이라 할 때, $e(n) \equiv d(n) - y_a(n)$, 순시자승오차 (instantaneous squared error) $\varepsilon^2(n)$은

$$\varepsilon^2(n) \equiv \frac{1}{2} e^2(n) = \frac{1}{2}(d(n) - y_a(n))^2 \tag{16}$$

으로 정의한다. 이러한 순시자승오차를 평균자승오차 (mean squared error)의 추정치로 하여 LMS 알고리즘을 적용하면, 중간층-출력층 연결 강도인 v_m $(m = 0, 1, \cdots, N_H)$을 위한 개선식 (update rule)은 아래와 같이 유도된다.

$$\begin{aligned} v_m(n+1) &= v_m(n) + \Delta v_m(n) \\ \Delta v_m(n) &= \eta e(n)\left(\sum_{k=0}^{N_I} a_{km}(n)\, x_k(n) + \beta_m(n)\right) \end{aligned} \tag{17}$$

여기서 η (> 0)는 학습 속도와 수렴에 관계된 학습률이다. $m \in M^0$인 경우

$$\mathrm{net}_m(n) = \sum_{k=0}^{N_I} w_{km}(n)\, x_k(n) \tag{18}$$

이므로, 결국 식 (18)은 다음과 같이 정리된다.

$$\Delta v_m(n) = \begin{cases} -\eta e(n) & \text{if } m \in M^- \\ +\eta e(n) & \text{if } m \in M^+ \\ \eta e(n)\{s_{i_m}(n)[\mathrm{net}_m(n) - u_{i_m-1}(n)] + \sigma(u_{i_m-1}(n))\} & \text{if } m \in M^0 \end{cases} \tag{19}$$

입력층-중간층 연결 강도인 w_{km}을 위한 개선식은 식 (15)와 (16)에 오차 역전파 과정을 적용하면 아래와 같이 유도된다.

$$\begin{aligned} w_{km}(n+1) &\equiv w_{km}(n) + \Delta w_{km}(n) \\ \Delta w_{km}(n) &= \begin{cases} 0 & \text{if } m \in M^- \\ 0 & \text{if } m \in M^+ \\ \eta e(n) v_m(n) s_{i_m}(n) x_k(n) & \text{if } m \in M^0 \end{cases} \end{aligned} \tag{20}$$

Ⅲ. 시뮬레이션 결과 및 고찰

다중계층 퍼셉트론 내의 sigmoid 활성함수에 대한 구간 선형 근사 방법과 기존의 양자화 방법의 성능을 비교하기 위해, 우리는 널리 benchmark 문제로 사용되는 XOR 문제에 대한 시뮬레이션을 수행하였다. 사용된 다중계층 퍼셉트론의 구조는 입력층, 은닉층, 출력층의 뉴론을 각각 2-2-1로 구성하고, 은닉층에서는 hyperbolic tangent 형태의 sigmoid 활성함수를 그리고 출력층에서는 선형 함수 $\sigma(u) = u$를 활성함수로 사용하였으며, 오차 역전파 알고리즘에서의 학습률 η는 0.1로 정하였다. XOR 문제에서 두 개의 입력은 -1 또는 +1의 값을 취하였고, 출력 역시 해당 입력쌍에 대해 -1 또는 +1의 값을 취하도록 하였다.

학습은 각 학습 epoch 마다 계산된 평균자승오차가 임계 평균자승오차보다 작으면 종료하며, 이 때 사용된 평균자승오차 $\overline{\varepsilon^2}$은 아래 식 (21)과 같이 정의된다.

$$\overline{\varepsilon^2} \equiv \frac{1}{L} \sum_{\ell=1}^{L} (d^{(\ell)} - y^{(\ell)})^2 \tag{21}$$

이 식에서 L은 학습 패턴의 개수 (본 실험의 경우 4), $d^{(\ell)}$과 $y^{(\ell)}$은 각각 ℓ ($\ell = 1, \cdots, L$)번째 패턴에 대한 목표 출력과 실제 신경회로망의 출력을 나타낸다. 본 실험에서는 임계 평균자승오차를 10^{-4}로 정하였다.

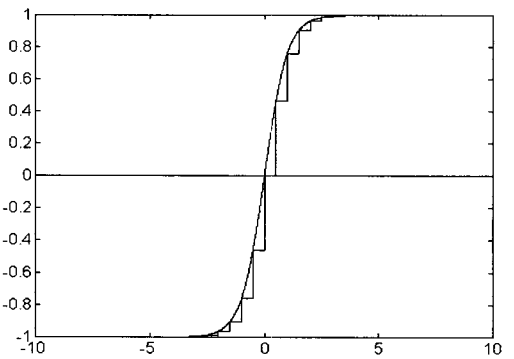


그림 4. look-up table을 이용한 sigmoid 활성함수의 양자화

Fig. 4. Quantization of sigmoid activation function using a look-up table.

구간 선형 근사 방법에서 은닉층 활성함수의 선형 근사를 위해 $u_0 = -5$, $u_{N-1} = 5$로 정하였으며, 그 안의 구간을 $N = 20$개의 구간으로 등간격 분할하였다. 양자화 방법을 이용하여 근사화된 sigmoid 활성함수는 그림 4와 같다. 이 그림의 활성함수도 구간 선형

근사한 경우 마찬가지로 [-5, 5]의 입력 범위를 Q 개의 등간격으로 분할하여 이에 따라 활성함수의 출력값을 양자화 하였으며, 그 외의 구간에 대해서는 활성함수의 입력이 $u_0=-5$보다 작은 경우는 -1로 정의하고 $u_{N-1}=5$보다 큰 경우는 1로 정의하였다.

본 실험에서는 구간 선형 근사 방법의 $N=20$인 경우와의 정확한 성능 비교를 위해 양자화 방법에서 100에서 10,000까지 Q 값을 변화시키면서, 각 Q 값에 따라 다른 초기 연결 강도를 이용하여 1,000번씩 Monte Carlo 실험을 수행하였다.

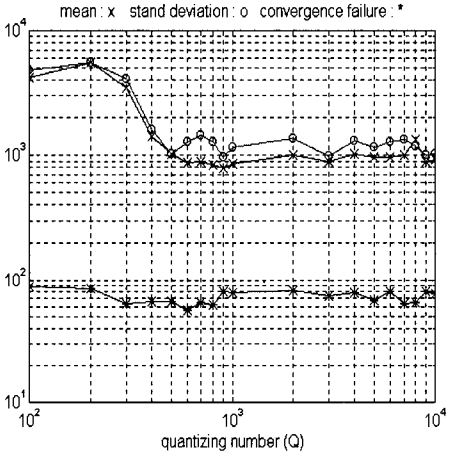


그림 5. 양자화 방법에서 $Q=100\sim10,000$에 따른 학습 수렴 성능 결과

Fig. 5. Learning convergence results according to $Q=100\sim10,000$ in quantization method.

그림 5와 표 1은 양자화 방법의 여러 Q 값마다 1,000번씩 Monte Carlo 실험을 수행한 결과로서, 학습이 성공한 경우 (즉 10^{-4} 보다 적은 평균자승오차)의 수렴 epoch 평균 횟수와 표준 편차, 그리고 1,000번의 실험 가운데 수렴 실패 횟수를 나타내고 있다. 본 연구에서는 25,000번의 epoch 내에 수렴하지 않은 경우를 "수렴 실패"로 정의하였다. 우리의 기초 연구 결과인 참고 문헌 [8]로부터 구간 선형 근사 방법의 $N=20$인 경우의 수렴 epoch 평균은 770, 수렴 epoch 표준편차는 864이고 수렴 실패가 하나도 없는 반면, 그림 5의 결과에서 양자화 방법이 구간 선형 근사 방법과 수렴 epoch 평균과 표준편차의 측면에서 비슷한 성능을 가지려면 Q = 1,000 정도가 되어야 함을 알 수 있다. 하지만 이런 경우에도 양자화 방법은 60회 이상의 수렴 실패가 발생한다. 결국 구간 선형 근사된 활성함수를 이용하는 경우 평균 수렴 횟수와 표준 편차를 비교할 때 LUT로 구현한 경우보다 나으며, 특히 수렴에 실패하는 경우가 LUT로 표현한 활성함수를 이용하는 경우보다 훨씬 적은 매우 우수한 성능을 보임을 알 수 있다. 이러한 결과는 적은 수의 구간을 갖는 본 제안 방법이 실제 디지털 하드웨어 구현 시 이미 언급한 바와 같이 수행 시간을 단축할 수 있을 뿐더러 안정적인 동작을 할 수 있음을 실험적으로 보인 것이라 할 수 있다.

아래의 표 2는 본 제안 방법과 양자화 방법에서 요구되는 저장량 (memory requirement), 탐색시간 (searching time), 그리고 계산량 (computational complexity)을 정리하여 보여준다. 여기서, 먼저 "저장량"은 $x_k\ (k=1,\cdots,N_I)$, $\text{net}_m\ (m=1,\cdots,N_H)$, $h_m\ (m=1,\cdots,N_H)$, y, $v_m\ (m=0,\cdots,N_H)$ 그리고 $w_{km}\ (k=0,\cdots,N_I,\ m=1,\cdots,N_H)$ 등과 같이 두 방법 모두 동일하게 요구되는 저장량을 제외하고 sigmoid 함수와 이의 미분값만을 위한 저장량을 의미한다. 다음 "탐색시간"은 이진 탐색 (binary search)을 가정하여 sigmoid 함수를 1회 탐색하는데 필요한 단계수를 의미하며, 마지막 "계산량"은 stochastic 학습 알고리즘을 적용 시 학습 과정의 일회 반복 당의 forward pass와 backward pass 각각에 필요한 곱셈수를 나타낸다. 이 계산량은 두 방법 모두 동일하게 sigmoid 함수의 포화 영역에 대하여 근사하여 이 부분에 대하여 간단히 계산될 수 있다는 사실을 무시한 "worst case" 결과이다.

표 2에서 고려한 양자화 방법은 크게 두 경우로 나눌 수 있다. 먼저 "sigmoid 미분을 저장하는 경우"는 앞서의 우리 논의대로 Q개의 sigmoid 함수값과 이 때의 미분값을 동시에 저장하는 경우이다. 다음 "sigmoid 미분을 저장하지 않는 경우"는 sigmoid 함수의 특성에 의해 이 함수의 미분값을 함수값으로 표현할 수 있다는 사실을 이용하는 경우이다. 즉, 만약 sigmoid 함수 $\sigma(u)$가 logistic 함수인 경우 이 함수의 미분 $\sigma'(u)$는 아래와 같이 표현된다.

$$\sigma'(u)=\sigma(u)\cdot(1-\sigma(u)) \tag{22}$$

만약 $\sigma(u)$가 hyperbolic tangent 함수인 경우

$$\sigma'(u)=1-\sigma^2(u) \tag{23}$$

과 같이 표현되므로, 실제 미분값을 저장하지 않고 함

표 1. XOR 문제에 대한 Monte Carlo 실험 결과 ($Q=100\sim1,000$)
Table 1. Monte Carlo simulation results for XOR problem ($Q=100\sim1,000$).

	양자화 방법										구간 선형 근사 방법
Look up table 크기 (Q)	100	200	300	400	500	600	700	800	900	1,000	N=20
수렴 epoch 평균	4,199	5,494	3,505	1,412	1,024	870	895	846	786	855	770
수렴 epoch 표준편차	4,844	5,589	4,086	1,599	1,024	1,290	1,440	1,277	971	1,172	864
1,000번 실험에서 수렴 실패 횟수	88	84	63	66	66	56	65	62	79	78	0

표 2. 저장량, 탐색시간, 계산량의 비교
Table 2. Comparisons of memory requirement, searching time and computational complexity.

	양자화 방법		구간 선형 근사 방법	본 논문의 XOR 문제에서 N=20, Q=1,000 인 경우 양자화 방법 대 구간 선형 근사 방법의 비	
	Sigmoid 미분을 저장하는 경우	Sigmoid 미분을 저장하지 않는 경우		양자화 방법에서 sigmoid 미분을 저장하는 경우	양자화 방법에서 sigmoid 미분을 저장하지 않는 경우
저장량	$2Q$	Q	$2N-1$	51.2	20.6
탐색시간	$\log_2 Q$	$\log_2 Q$	$\log_2(N-1)$	2.35	2.35
계산량	Forward : $(N_I+1)\cdot N_H$ Backward : $(4N_I+2)\cdot N_H$	Forward : $(N_I+1)\cdot N_H$ Backward : $(4N_I+3)\cdot N_H$	Forward : $(N_I+2)\cdot N_H$ Backward : $(4N_I+2)\cdot N_H$	0.93	1.00

수값을 이용하여 미분을 구할 수 있다. 따라서, 이 경우 양자화 방법의 소요 저장량은 반으로 줄게 되나, backward pass에서 각 미분값을 구하기 위해 하나의 은닉층 뉴론 당 1회의 곱셈이 더 소요된다.

또한 위 표에서 저장량의 경우 본 제안 방법은 N 개의 구간 경계값과 $N-1$개 구간에서의 미분값을 저장한다. 계산량의 경우 본 제안 방법이 forward pass 에서 N_H번이 추가로 필요한 이유는 net_m이 계산된 후 식 (7)에 의해 $\sigma_a(\text{net}_m)$이 계산될 때 매 중간층마다 1회의 추가적인 곱셈이 필요하기 때문이다. 위 표로부터 우리는 본 제안 방법이 양자화 방법에 비해 약간의 계산량 증가가 필요하나 (혹은 동일한 계산량) 큰 저장량 및 탐색시간의 감소를 얻을 수 있음을 알 수 있다.

본 논문에서 고려한 기존의 "양자화" 방법은 다수의 sigmoid 함수값을 LUT에 저장한다는 의미이며, 실제 모의 실험 시 이들 값을 표현하는 정밀도 (precision)는 제안된 방법에서 구간 경계값들을 표현하는 정밀도와 동일하다는 점을 특기할 만 하다. 본 제안 방법을 실제 하드웨어로 구현할 경우 아날로그의 구간 경계값들을 그대로 사용하는 것이 아니고 디지털로 재차 변환 (일반적인 의미의 양자화)하여 저장하여야 하며, 이는 본 논문에서 고려하는 기존의 "양자화" 방법의 함수값들에게도 역시 동일하게 적용된다. 즉, 본 논문에서 기존의 "양자화" 방법은 실제 하드웨어에 디지털로 변환하여 저장하는 양자화 과정을 의미하는 것이 아니며, 그 이전의 알고리즘 단계에서의 방법을 의미한다.

Ⅳ. 결 론

본 논문에서는 다중계층 퍼셉트론을 디지털 하드웨어로 구현하는 경우 이 신경회로망 모델 내의 비선형 sigmoid 활성함수를 구현하는 방법으로 널리 사용되는 look-up table (LUT)에 의한 양자화 방법의 단

점 (적절한 수준의 양자화 오차를 위해 많은 량의 저장이 필요)을 제거할 수 있는 새로운 방법으로서 비선형 sigmoid 활성함수의 입력 범위를 (적은 수의) 여러 구간으로 나누고 각 구간별로 선형 근사하는 "구간 선형 근사 (piecewise affine approximation)" 방법을 제안하고, 이를 이용한 다중계층 퍼셉트론의 표현식과 오차 역전파 학습 알고리즘을 유도하였으며, XOR 문제에 대한 Monte Carlo 시뮬레이션을 통해 성능을 비교하였다. 실험 결과, 적은 수의 구간 ($N=20$)을 갖는 구간 선형 근사 방법이 평균 학습 수렴 횟수와 표준 편차를 비교할 때 매우 많은 량의 양자화 레벨 (Q = 800 ~ 1,000)을 갖는 양자화 방법보다 나으며, 특히 수렴에 실패하는 경우가 LUT로 표현한 활성함수를 이용하는 경우보다 훨씬 적은 매우 우수한 성능을 보임을 알 수 있다. 또한 본 제안 방법은 약간의 계산량 증가가 필요하나 (혹은 동일한 계산량) 큰 저장량 및 탐색시간의 감소를 얻을 수 있음 역시 알 수 있었다. 수학적인 함수 근사 능력의 측면을 볼 때, 본 연구에서 제안하는 구간 선형 근사된 sigmoid 활성함수를 사용하는 다중계층 퍼셉트론은 원래의 sigmoid 활성함수를 사용하는 경우와 마찬가지로 우수한 수학적 함수 근사 능력을 그대로 보존함을 알 수 있으며, 이는 구간 선형 근사된 sigmoid 활성함수가 참고문헌 [3] - [5] 에서 정의한 일반적인 sigmoid 함수군 (函數群)에 포함되기 때문이다.

실제 sigmoid 함수를 선형 근사하여 사용하는 방법은 널리 논의되어 왔다 (예를 들어, 참고문헌 [9] 의 Figure 1.7). 하지만 이러한 논의에서는 미분이 불가능한 signum 함수[9]와 미분 가능한 sigmoid 함수 사이의 중간 형태로서 이를 간단하게 기술하는 경우가 대부분이며, 구간 선형 근사 함수의 형태도 단순히 하나의 선형 구간만을 갖는 경우를 주로 고려하고 있다. 따라서 구간 선형 근사 방법을 양자화 방법을 대신하는 디지털 하드웨어 구현을 위한 효과적인 방법으로 인식하고, 임의의 개수와 간격을 허용하는 구간을 고려하여 이를 사용하는 다중계층 퍼셉트론에 대한 해석과 학습 알고리즘의 유도, 그리고 양자화 방법과의 실험적 비교를 종합적으로 수행한 본 논문은 이러한 일반적인 논의와 큰 차이를 갖는다 할 수 있다.

앞으로의 연구 과제로서, 먼저 본 논문의 결과를 확장하여 여러 개의 은닉층, logistic 활성함수, 여러 개의 출력층 뉴론 등과 같은 일반적인 형태를 갖는 다중계층 퍼셉트론에 구간 선형 근사 방법을 적용하여, 이를 해석하고 이에 적합한 오차 역전파 학습 알고리즘을 유도하여야 한다. 또한, 구간 선형 근사 방법과 양자화 방법에 의한 오차를 이론적으로 해석·비교하는 작업이 수행되어야 한다. 실제 신경회로망 내의 연결 강도나 입력을 양자화 할 경우에 발생하는 오차에 대한 통계적 해석은 많이 보고되고 있으나[10]-[13], 활성함수의 양자화나 선형 근사에 의한 오차에 대한 해석은 매우 드문 것이 현실이다. 하지만 실제 신경회로망의 디지털 하드웨어 구현에서는 이러한 측면을 반드시 고려하여야 하며, 따라서 본 제안 방법에 의한 오차에 대한 이론적 해석뿐만이 아니고 양자화 방법의 경우에 대한 해석 역시 매우 필요하다 할 수 있다. 이러한 수학적인 결과들은 실제 디지털 하드웨어 구현 시 사용자가 미리 성능을 예측, 분석하는데 매우 유용하게 사용될 수 있을 것이다. 마지막으로, 본 제안 방법과 양자화 방법을 우리가 실제 접하는 응용 문제에 적용하여 본 제안 방법의 효용성을 입증하는 작업 역시 필요하다 할 수 있다.

참 고 문 헌

[1] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*, Prentice-Hall, 1985.

[2] J. McClelland and D. Rumelhart, *Parallel Distributed Processing*, vol. 1, The MIT Press, 1987.

[3] K. Hornik, M. Stinchcombe and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, pp. 359-366, 1989.

[4] K. Funahashi, "On the approximate realization of continuous mappings by neural networks," *Neural Networks,* vol. 2, pp. 183-192, 1989.

[5] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Contr., Signal and Syst.,* vol. 2, pp. 303-314, 1989.

[6] N. Morgan (Ed.), *Artificial Neural Networks Electronic Implementations*, IEEE Computer Society Press, 1990.

[7] S. Anna Durai *et al.*, "A learning strategy for multilayer neural network using

discretized sigmoidal function," *Proc. IEEE Int'l Conf. Neural Networks,* Vol. 4, pp. 2107-2110, Perth, West Australia, November, 1995.

[8] 윤병문, 신요안, "다중계층 퍼셉트론에서 sigmoid 활성함수의 구간 선형 근사," *1996년도 대한전자공학회 하계종합학술대회 논문집,* pp. 707-710, 1996년 6월

[9] S. Haykin, *Neural Networks-A Comprehensive Foundation,* IEEE Press, 1994.

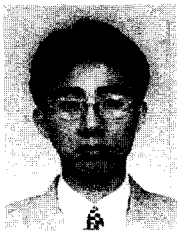
[10] J. Y. Choi and C. H. Choi, "Sensitivity of multilayer perceptron with differentiable activation functions," *IEEE Trans. on Neural Networks*, vol. 3, no. 1, pp. 101-107, January 1992.

[11] S.-H. Oh and Y. Lee, "Sensitivity analysis of single hidden-layer neural networks with threshold functions," *IEEE Trans. on Neural Networks*, vol. 6, no. 4, pp. 1005-1007, July 1995.

[12] G. Dundar and K. Rose, "The effect of quantization on multilayer neural networks," *IEEE Trans. on Neural Networks*, vol. 6, no. 6, pp. 1446-1451, November 1995.

[13] N. S. Merchawi *et al.*, "A probabilistic model for the fault tolerance of multilayer perceptrons," *IEEE Trans. on Neural Networks*, vol. 7, no. 1, pp. 201-205, January 1996.

저 자 소 개



尹 炳 文(正會員)

1969년 3월 24일생. 1995년 2월 숭실대학교 공과대학 전자공학과(공학사). 1997년 2월 숭실대학교 대학원 전자공학과(공학석사). 1997년 3월 ~ 현재 LG정보통신(주) 중앙연구소 이동통신연구단 연구원. 관심분야는 디지털 통신 시스템, 신경회로망 응용

辛 堯 安(正會員)

1965년 1월 19일생. 1987년 2월 서울대학교 공과대학 전자공학과(공학사). 1989년 2월 서울대학교 대학원 전자공학과(공학석사). 1992년 12월 University of Texas at Austin 전기 및 컴퓨터공학과(Ph.D.). 1992년 12월 ~ 1994년 7월 Austin 소재 MCC(Microelectronics & Computer Technology Corp.) 연구 콘소시엄 연구원. 1994년 9월 ~ 현재 숭실대학교 전자공학과 조교수. 1996년 3월 ~ 현재 한국퍼지및지능시스템학회 논문지 편집위원. 관심분야는 이동통신 시스템, 통신 신호 처리, 신경회로망의 통신 응용