하이브리드 VLSI 신경망 프로세서에서의 양자화에 따른 영향 분석

권 오 준† · 김 성 우† · 이 종 민†

요 약

인공 신경망을 실제적인 응용 분야에 적용하기 위하여 하드웨어 시스템으로 구현하는 것이 필요하다. 하드웨어로 구현하는 방법에는 현재 하이브리드 VLSI 신경망 칩으로 구현하는 것이 가장 유망하다. 이미 학습된 신경망을 하이브리드 신경망 칩을 사용하여 구현하는 경우 뉴런 출력과 가중치 값의 양자화 과정이 필수적이다. 이러한 과정은 신경망의 출력층 뉴런의 이미 학습된 출력에 비해 왜곡을 야기한다. 본 논문에서는 이러한 신경망의 출력 왜곡에 대한 통계적 특성을 자세하게 분석하였다. 분석 결과는 신경망의 출력 왜곡을 줄이기 위해서는 입력 벡터의 정규화와 가중치 값들이 작아야 한다는 사실을 보여 주었다. 시계열 데이터에 대한 실험 결과는 분석 결과를 고려하여 학습된 신경망들의 경우 실제로 뉴런 출력 및 가중치 값의 양자화로 인한 출력층 뉴런의 출력 왜곡이 상당히 줄어들 수 있음을 명확히 보여 주었다.

Analysis of the Effect on the Quantization of the Network's Outputs in the Neural Processor by the Implementation of Hybrid VLSI

Oh-Jun Kwon† · Seong-Woo Kim† · Jong-Min Lee†

ABSTRACT

In order to apply the artificial neural network to the practical application, it is needed to implement it with the hardware system. It is most promising to make it with the hybrid VLSI among various possible technologies. When we implement a trained network into the hybrid neuro-chips, it is to be performed the process of the quantization on its neuron outputs and its weights. Unfortunately this process cause the network's outputs to be distorted from the original trained outputs. In this paper we analysed in detail the statistical characteristics of the distortion. The analysis implies that the network is to be trained using the normalized input patterns and finally into the solution with the small weights to reduce the distortion of the network's outputs. We performed the expriment on an application in the time series prediction area to investigate the effectiveness of the results of the analysis. The experiment showed that the network by our method has more smaller distortion compared with the regular network.

키워드 : 신경망 칩(Neuro-chip), 하드웨어(Hardware), 출력 왜곡(Distortion of the network's outputs)

1. 서 론

인간 두뇌의 정보처리 메카니즘으로부터 영감을 받은 인공 신경망이 활발한 연구의 결과로 패턴 인식, 최적화 문제, 로봇 제어 및 시계열 예측 등 기존의 디지털 컴퓨터가 해결하기 어려웠던 분야에서 좋은 결과를 보고하고 있다[1]. 이러한 연구 결과에 힘입어 많은 연구자들이 인공 신경망의 대규모 병렬 처리의 장점을 극대화하기 위하여 하드웨어로 구현하고자 하는 많은 노력을 하여 왔다. 현재 구현의 용이성, 빠른 처리 속도, 대규모 신경망의 구성이 용이하다는 장점으로 인해 VLSI 신경망 칩을 이용한 방법이 많이 연구되고 있다[2]. VLSI 신경망 칩을 제조하는 방법에는 크게 아날로그 방식과 디지털 방식이 있다. 최근에는 이들 두 방식의 장점을 혼합한 하이브리드 방식이 많이 사용되고 있다[3]. 한편 디지털 또는 하이브리드 방식으로 설계된 신경망 칩을 사용하여 신경망을 하드웨어로 구현하는 경우 뉴런의 출력과 가중치의 양자화가 필수적이다. 이로 인해 이미 학습된 신경망을 하이브리드 신경망 칩을 사용하여 하드웨어로 구현할 때, 각 뉴런들의 출력에서 이미 학습된 출력에 비해 왜곡 현상이 나타날 수 있는 문제가 발생한다. 따라서 인공신경망의 실용적인 응용 연구를 위해서는 이러한 양자화가 이미 학습된 신경망의 성능에 어떠한 영향을 미치는 가에 대한 자세한 영향 분석이 해결해야 할 중요한 문제이다.

이제까지 많은 연구자들이 각각 서로 다른 뉴런 모델과 여러 가지 가정을 함으로써 신경망 구성요소에 대한 잡음이나 양자화로 인한 신경망 출력에 미치는 영향을 분석하려고 시도하였다. Stevenson과 Widrow[4]는 기하학적인 분석 방법을 사용하여 입력 패턴과 가중치의 값들의 분포를 단순히 일양 분포로 가정하고 가중치의 작은 잡음에 대한 출력층 뉴런에서의 출력 민감도를 분석하였다. 그리고 가중치의 변화율에 대한 출력의 전도 확률을 유도하였다. 이 연

† 정 회 원 : 동의대학교 컴퓨터 · 영상공학부 교수
논문접수 : 2001년 7월 23일, 심사완료 : 2002년 7월 6일

구의 분석 결과로 뉴런당 가중치의 수와 계층당 뉴런의 수는 양자화로 인한 영향과는 무관함을 주장하였다. Xie와 Jabri[5]는 양자화 에러를 정상 랜덤 과정인 백색 잡음으로 그리고 입력 패턴, 가중치 및 뉴런의 입력인 가중치 합을 모두 일양 분포로 가정하고 뉴런 출력에서의 잡음대 신호비를 유도하였다. 분석 결과에 따르면 3층 퍼셉트론 신경망의 경우 은닉층과 출력층 사이의 가중치에 비해서 입력층과 은닉층 사이의 가중치의 영향이 더 크게 나타남을 보였다. 그러나 이들의 연구들은 둘 다 뉴런의 출력이 단순히 +1, -1 두 값만을 가지는 뉴런들로 구성된 다층 퍼셉트론 신경망인 madaline을 대상으로 하여 대상 모델이 너무 단순하다는 단점이 있다. Dündar와 Rose[6]는 앞서 Xie와 Jabri가 분석한 모델에서 뉴런의 활성화 함수를 일반적으로 많이 사용되는 시그모이드 형태를 가진 로지스틱 함수로 변경함으로써 좀 더 현실적인 모델을 분석 대상으로 하였다. 이들은 가중치의 정밀도에 따른 출력층 뉴런의 출력에서의 신호대 잡음비를 유도하였다. 또한 이들은 A/D 변환 문제에 대한 실험을 통하여 자신들의 결과가 Xie와 Jabri에 의한 결과보다 더 정확함을 증명해 보였다. 그러나 이들은 분석을 간략히 하기 위해 은닉층 뉴런의 출력을 일양분포로 가정함에 따라 비현실적이다.

Piché[9]는 이미 학습된 신경망보다는 랜덤한 입력 패턴과 가중치를 가지는 신경망에 대해 입력 또는 가중치의 작은 에러에 대한 출력층 뉴런에서의 출력 분산을 분석하였다. 이 분석 결과를 이용하여 가중치 에러에 대한 출력층 뉴런에서의 잡음대 신호비를 유도하였다. 이러한 유도 결과를 이용하여 대상 신경망의 허용 잡음대 신호비가 주어지면 이를 만족하는 가중치의 정밀도를 계산할 수 있는 방법을 제시하였다. Choi와 Choi[7]는 입력 또는 가중치의 잡음을 백색 잡음으로 가정하고 이들이 각각 덧셈적인 또는 곱셈적인 잡음인 경우에 대해 출력층 뉴런에서의 출력 민감도를 다음과 같이 정의하고 이를 유도하였다.

$$S(\mathbf{w}) = \lim_{\sigma \to 0} \frac{var[\Delta \mathbf{x}(L)]}{\sigma}$$

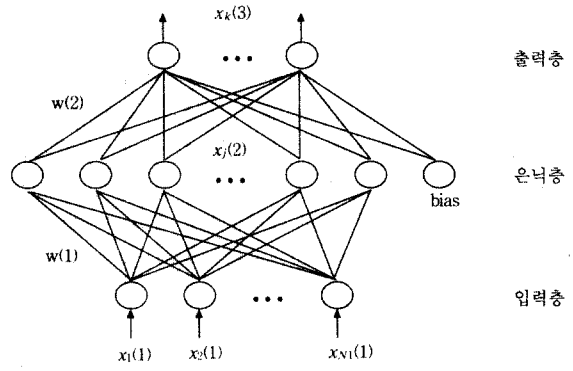
위 식에서 σ는 가중치 변화량의 표준 편차, $\Delta \mathbf{x}(L)$는 출력층 뉴런에서의 출력 에러, $var[\cdot]$은 분산을 나타낸다. 그리고 잡음에 대한 민감도를 줄이는 방안의 하나로 먼저 여러개의 신경망을 학습시킨 다음 이들에 대해 각각의 출력 민감도를 계산하여, 가장 낮은 민감도를 가진 신경망을 선택하는 방법을 제시하였다. Lovell[8] 등은 입력 또는 가중치의 작은 잡음에 대한 출력층 뉴런에서의 출력 에러와 분산의 상한을 유도하였다. 이들 연구는 비교적으로 현실적인 모델 및 가정하에 분석하였으나 강한 비선형 특성을 가지고 있는 뉴런의 출력을 선형 근사하는 등 분석 과정에서 많은 근사를 통하여 단순화하였고 제시된 대책이 소극적이다.

한편 다층 퍼셉트론 신경망은 강한 비선형 특성을 가진 많은 뉴런들이 계층적으로 연결되어 동작하기 때문에 뉴런 출력 및 가중치의 양자화로 인한 작은 에러는 다음층 뉴런들로 전달되면서 점차적으로 증폭된다. 그리고 신경망을 하드웨어로 구현하여 실용적인 문제에 적용할 경우 대부분 신경망의 크기가 충분히 클 것이다. 그러므로 신경망의 내부 구성 요소들에서 발생하는 에러는 그것이 작은 에러라 할 지라도 최종 출력층 뉴런에서의 출력 에러는 무시할 수 없을 정도로 크게 증폭될 수 있다. 따라서 다층 퍼셉트론 신경망의 뉴런 출력과 가중치의 양자화로 인한 영향을 최소화하기 위한 방안을 찾기 위해서는 먼저 가능한 한 그러한 영향을 자세하게 분석하여야 한다.

본 논문에서는 다층 퍼셉트론 신경망에서 뉴런 출력과 가중치의 양자화로 인하여 발생하는 출력층 뉴런에서의 출력 왜곡에 대한 통계적인 특성을 뉴런의 비선형 전달 함수에 대한 이차 근사까지 이용함으로써 보다 자세히 분석한다. 그리고 이미 학습된 특정 신경망만을 대상으로 분석하기 보다는 동일한 문제에 대해 서로 다른 가중치들로 초기화를 통해 학습된 일련의 다층 퍼셉트론 신경망들의 모임(ensemble)을 대상으로 한다. 기존의 대부분 연구들이 분석 대상 모델을 너무 간략화하였거나 비현실적인 가정을 한 반면 본 논문에서는 어떠한 비현실적인 가정도 하지 않고 실제 가장 많이 사용되는 실용적인 모델에 대하여 분석한다. 그리고 이러한 분석 결과로 부터 양자화로 인한 출력 왜곡을 줄일 수 있는 방안에 대한 고찰을 하고 시계열 데이터에 대한 실험을 통하여 그 결과를 확인한다.

2. 다층 퍼셉트론 신경망

인공 신경망은 단위 뉴런들을 연결하는 구조와 각 뉴런들이 상호 작용하는 방법에 따라 지금까지 많은 모델이 고안되고 연구되어 왔다. 이러한 모델들 중에서 현재 많은 응용 분야에서 가장 널리 사용되고 있으며 본 논문에서 다루게 될 다층 퍼셉트론 신경망 모델을 설명하고 앞으로 사용하게 될 용어와 기호들을 정의한다.



(그림 1) 3층 퍼셉트론 신경망

다층 퍼셉트론 신경망은 3개 이상의 층으로 구성되며 각 층은 여러개의 뉴런들로 구성된다. 여기서는 L개의 층으로

구성된 일반적인 다층 퍼셉트론 신경망을 고려한다. 각 층을 1에서 L로 인덱스한다. 첫 번째 층은 입력층이고, 마지막 층인 L번째 층은 출력층이다. 그리고 입력층과 출력층사이에 있는 두 번째 층부터 $(L-1)$번째 층까지는 모두 은닉층이다. k번째 층의 뉴런의 수는 N_k로 나타낸다. k번째 층의 i번째 뉴런의 출력은 $x_i(k)$로 나타내고 k번째 층의 전체 뉴런들의 출력 벡터를 $\mathbf{x}(k)$로 나타낸다. (그림 1)은 하나의 은닉층만을 가진 다층 퍼셉트론 신경망의 예를 보여준다. 이렇게 입력층, 은닉층, 출력층의 3층만으로 구성된 신경망을 **3층 퍼셉트론 신경망**이라 부른다. 3층 퍼셉트론 신경망에 대해서 언급할 경우에는 이해를 쉽게 하기 위하여 단순히 입력 패턴 벡터를 $\mathbf{x}$, 은닉층 뉴런 출력 벡터를 $\mathbf{z}$, 출력층 뉴런의 출력 벡터를 $\mathbf{y}$라 나타낸다. 그리고 k번째 층에 있는 모든 뉴런들은 단지 $(k+1)$번째 층에 있는 뉴런들과만 순방향 연결이 있다. 이들 연결선들의 가중치는 $N_k \times N_{k+1}$ 차원의 행렬 $\mathbf{w}(k)$로 표현한다. k번째 층의 j번째 뉴런과 $(k+1)$번째 층의 i번째 뉴런을 연결하는 가중치는 $w_{ij}(k)$로 나타내고, 신경망 전체의 가중치 집합을 ω로 나타낸다. 그리고 $(k-1)$번째 층으로 부터 k번째 층으로의 순방향 신호 전달식은 다음과 같다.

$$x_i(k) = T(a_i(k)) \tag{1}$$

위 식에서

$$a_i(k) = \sum_{j=0}^{N_{k-1}} w_{ij}(k-1)x_j(k-1) \tag{2}$$

$T(\cdot)$는 뉴런 활성화 함수로 본 논문에서는 일반적으로 가장 널리 사용되고 있는 다음과 같은 시그모이드 형태의 로지스틱 함수를 사용한다.

$$T(z) = \frac{1}{1+e^{-gz}} \tag{3}$$

3. 분석 및 고찰

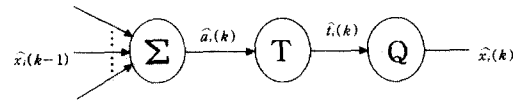
먼저 신경망의 뉴런 출력 및 가중치 값의 양자화 과정에 대해 설명한다. 일반적으로 양자화하고자 하는 수, u의 범위가 [-A, A]이고, 양자화 정밀도가 n비트로 주어지면 양자화 단계 크기, q는 다음과 같이 결정된다.

$$q = \frac{A}{2^{n-1}-1} \tag{4}$$

양자화는 식 (5)와 같이 정의되는 양자화 함수, $Q[\cdot]$에 의해 수행되고, 양자화 후의 값은 양자화 단계 크기 q의 정수배에 해당하는 값들을 갖는다.

$$Q[u] = mq \quad \text{if } (m-\frac{1}{2})q \le u < (m+\frac{1}{2})q \tag{5}$$

여기서 m은 정수이다. 앞으로의 분석 과정에서 양자화가 수행된 후의 k번째 층의 i번째 뉴런의 입출력에 대한 부호는 (그림 2)를 따른다. (그림 2)에서 '**T**'는 뉴런의 활성화 함수, '**Q**'는 양자화 함수를 의미한다. 즉, 뉴런 입력은 $\hat{a}_i(k)$, 양자화 수행 전의 출력은 $\hat{t}_i(k)$, 양자화 수행후의 출력은 $\hat{x}_i(k)$로 나타낸다.



(그림 2) 양자화된 신경망에서 뉴런의 입출력 신호들

한편 뉴런 출력과 가중치의 양자화에 대한 정밀도 요구사항은 다를 수 있다[9]. 이에 따라 앞으로 뉴런의 출력 및 가중치에 대한 양자화 비트 수를 각각 b_n과 b_w으로 나타낸다. 그리고 신경망 칩을 사용하여 인공신경망을 하드웨어로 구현하였을 때 신경망의 입력은 입력을 받아들이는 센서로부터 입력층 뉴런의 입력으로 직접 전달된다. 즉 입력층 뉴런에서는 어떠한 변환도 거치지 않기 때문에 입력층의 뉴런 출력에 대해서는 양자화를 하지 않는다.

3.1 첫 번째 은닉층에서의 출력 왜곡 특성 분석

양자화된 신경망에서 주어진 입력 벡터에 대한 첫 번째 은닉층의 i번째 뉴런 입력, $\hat{a}_i(2)$는 다음과 같다.

$$\begin{aligned}\hat{a}_i(2) &= \sum_{j=1}^{N_1} (w_{ij}(1) + \Delta w_{ij}(1))\, x_j(1) \\ &= a_i(2) + \sum_{j=1}^{N_1} \Delta w_{ij}(1) x_j(1)\end{aligned} \tag{6}$$

위 식에서 $\Delta w_{ij}(1)$은 원래 학습된 가중치 $w_{ij}(1)$에 대한 양자화 에러이다. 첫 번째 은닉층 뉴런의 입력 왜곡, $\Delta a_i(2)$는

$$\Delta a_i(2) = \sum_{j=1}^{N_1} \Delta w_{ij}(1) x_j(1) \tag{7}$$

$\Delta w_{ij}(1)$은 양자화 에러이므로 구간 $[-q_w/2,\ q_w/2]$에서 일양 분포을 따르고, 평균이 0이고 분산 $(\sigma^2_{\Delta w})$이 $q_w^2/12$인 랜덤 확률 변수이다. 여기서 q_w는 가중치 값의 범위 $[-W_{\max},\ W_{\max}]$와 양자화 정밀도 b_w에 대해 식 (4)에 의해 결정되는 양자화 단계 크기이다. 한편 신경망을 하드웨어로 구현하여 실용적인 문제에 적용하는 경우 대부분 신경망의 크기가 충분히 클 것이다. 입력의 차원이 통계적으로 충분히 크다고 가정하면 중심 극한 정리에 의해 $\Delta a_i(2)$를 정규 분포를 따르는 확률 변수로 취급할 수 있다[10]. 그러면 $\Delta a_i(2)$의 평균과 분산을 각각 다음과 같이 얻을 수 있다.

$$\begin{aligned}E[\Delta a_i(2)] &= E[\sum_{j=1}^{N_1} \Delta w_{ij}(1) x_j(1)] \\ &= 0\end{aligned} \tag{8}$$

$$Var[\Delta a_i(2)] = E[\Delta^2 a_i(2)] - E^2[\Delta a_i(2)] = \sigma^2_{\Delta w} \sum_{j=1}^{N_1} x_j^2(1) \quad (9)$$

위 식의 유도과정에서 가중치 에러 및 뉴런 입력의 에러에 대한 기대치가 0이고, 가중치들의 에러들은 서로 독립이라는 사실을 이용하였다. 한편 원래의 뉴런 입력과 그것의 에러의 곱에 대한 기대치는

$$E[a_i(2)\Delta a_i(2)] = E[\sum_{j=1}^{N_1} w_{ij}(1)x_j(1) \sum_{j=1}^{N_1} \Delta w_{ij}(1)x_j(1)] = 0 \quad (10)$$

따라서 뉴런 입력과 그것의 에러에 대한 공분산

$$Cov[a_i(2), \Delta a_i(2)] = 0 \quad (11)$$

앞의 결과들을 이용하여 정리하면, 양자화 후의 뉴런 입력, $\hat{a}_i(2)$는 정규 분포를 따르며 그것의 기대치와 분산은 다음과 같이 얻어진다.

$$E[\hat{a}_i(2)] = a_i(2) \quad (12)$$

$$Var[\hat{a}_i(2)] = E[\hat{a}_{i^2}(2)] - E^2[\hat{a}_i(2)] = Var[\Delta a_i(2)] \quad (13)$$

$$\hat{a}_i(2) \sim n(\mu_{a_i(2)}, \sigma_{a_i(2)}) \quad (14)$$

위 식에서 $\mu_{a_i(2)} = E[\hat{a}_i(2)]$이고, $\sigma_{a_i(2)} = \sqrt{Var[\hat{a}_i(2)]}$ 이다.

이제 양자화된 신경망의 뉴런 입력 $\hat{a}_i(2)$는 뉴런의 비선형 활성화 함수를 거치게 된다. 뉴런의 출력은 뉴런의 활성화 함수 $T(\cdot)$에 대하여 Taylor series 확장을 적용하여 중심 $a_i(2)$에서 이차 근사까지 취함으로써 다음과 같이 근사할 수 있다.

$$\hat{t}_i(2) = T(a_i(2) + \Delta a_i(2)) \cong T(a_i(2)) + \Delta a_i(2)\, T'(a_i(2)) + \frac{\Delta a_i^2(2)}{2} T''(a_i(2)) \quad (15)$$

위의 근사식으로 부터 첫 번째 은닉층 뉴런의 출력에 대한 기대값과 분산을 다음과 같이 얻을 수 있다.

$$E[\hat{t}_i(2)] = T(a_i(2)) + \frac{T''(a_i(2))}{2} Var[\hat{a}_i(2)] \quad (16)$$

$$Var[\hat{t}_i(2)] = E[\hat{t}_i^{\,2}(2)] - E^2[\hat{t}_i(2)] \cong [T_1'(a_i(2))]^2 Var[\hat{a}_i(2)] \quad (17)$$

위 식의 유도 과정에서 $E[\Delta^3 a_i(2)] = 0$ 라는 사실을 이용하였다. 그리고 가중치 정밀도(b_w)에 비해 가중치 값의 범위 $[-W_{\max}, W_{\max}]$가 너무 크지 않다면 $q_w \ll 1$ 이므로 뉴런 입력 왜곡의 더 이상의 고차항에 대한 모멘트는 무시하였다.

마지막으로 첫 번째 은닉층 뉴런의 출력에 대한 양자화 에러를 고려해야 한다. 입력층과 첫 번째 은닉층 사이의 가중치 값 양자화로 인해 전파된 에러를 고려하지 않고, 단지 첫 번째 은닉층 뉴런의 출력에 대한 양자화로 인한 에러 Δx는 구간 $[-q_x/2, q_x/2]$에서 일양 분포를 따른다. 따라서 Δx는 평균이 0이고 분산($\sigma^2_{\Delta x}$)이 $q_x^2/12$인 랜덤 확률변수이다. 여기서 q_x는 뉴런 출력의 양자화 단계 크기이다. 뉴런의 활성화 함수가 로지스틱 함수이므로 은닉층 뉴런의 출력은 [0, 1]의 연속적인 값을 갖는다. 따라서 q_x는 식 (4)에 의해 다음과 같이 결정된다.

$$q_x = \frac{1}{2^{b_w} - 1} \quad (18)$$

이 양자화 에러는 뉴런 출력과는 독립이고 덧셈적인 에러로 취급이 된다.

양자화 후의 첫 번째 은닉층 뉴런에서의 최종 출력 왜곡, $\Delta x_i(2)$는

$$\Delta x_i(2) = \hat{x}_i(2) - x_i(2) \quad (19)$$

위 식에서

$$\hat{x}_i(2) = Q[\hat{t}_i(2)] \quad (20)$$

마지막으로 $\Delta x_i(2)$에 대한 기대값과 분산은 다음과 같다.

$$E[\Delta x_i(2)] = \frac{T''(a_i(2))}{2} Var[\Delta a_i(2)] \quad (21)$$

$$Var[\Delta x_i(2)] = [T'(a_i(2))]^2 Var[\Delta a_i(2)] + \sigma^2_{\Delta x} \quad (22)$$

3.2 임의의 은닉층에서의 출력 왜곡 특성 분석

양자화된 신경망에서 $k(\geq 3)$번째 층의 뉴런 입력 왜곡은 그 이전의 모든 층에서 전파된 에러의 영향을 받은 $(k-1)$번째 층의 뉴런 출력 왜곡과 $(k-1)$번째 층과 k번째 층을 연결하는 가중치 값들의 양자화로 인한 에러들에 의해 야기된다. 따라서 k번째 층의 i번째 뉴런 입력, $\hat{a}_i(k)$는 다음과 같다.

$$\hat{a}_i(k) = \sum_{j=1}^{N_{k-1}} (w_{ij}(k-1) + \Delta w_{ij}(k-1))(x_j(k-1) + \Delta x_j(k-1)) \cong a_i(k) + \sum_{j=1}^{N_{k-1}} w_{ij}(k-1)\Delta x_j(k-1) + \sum_{j=1}^{N_{k-1}} \Delta w_{ij}(k-1)x_j(k-1) \quad (23)$$

위 식에서 $\Delta x_j(k-1)$은 이전의 모든 층으로 부터 전파된 에러의 영향으로 결과적으로 나타나는 $(k-1)$번째 층의 j번째 뉴런 출력 왜곡이다. 위의 식 (23)에서 가중치 에러 $\Delta w_{ij}(k-1)$과 이전층 뉴런에서의 출력 왜곡 $\Delta x_j(k-1)$의 곱에 의해 생성되는 에러는 매우 작기 때문에 무시하였다. 식 (23)의 오른편 두 번째와 세 번째 항을 각각 Δ_1, Δ_2로 둔다.

$$\Delta a_i(k) \cong \Delta_1 + \Delta_2 \tag{24}$$

Δ_1은 이전층인 $(k-1)$번째 층 뉴런들의 출력 왜곡에 의해 발생되는 에러이고, Δ_2는 $(k-1)$번째 층과 k번째 층을 연결하는 가중치 값들의 양자화로 인하여 발생되는 에러이다. 앞에서와 마찬가지로 $\Delta w_{ij}(k-1)$는 구간 $[-q_w/2, q_w/2]$에서 일양 분포를 가지므로 평균이 0이고 분산 $(\sigma^2_{\Delta w})$이 $q_w^2/12$인 랜덤 확률 변수이다. 따라서 Δ_1과 Δ_2 각각에 대한 기대값과 분산은 다음과 같이 얻을 수 있다.

$$E[\Delta_1] = \sum_{j=1}^{N_{k-1}} w_{ij}(k-1)\, E[\Delta x_j(k-1)] \tag{25}$$

$$Var[\Delta_1] = \sum_{j=1}^{N_{k-1}} w_{ij}^2(k-1)\, Var[\Delta x_j(k-1)] \tag{26}$$

$$E[\Delta_2] = 0 \tag{27}$$

$$Var[\Delta_2] = \sigma^2_{\Delta w} \sum_{j=1}^{N_{k-1}} x_j^2(k-1) \tag{28}$$

그리고 에러의 근원이 되는 $(k-1)$번째 층의 뉴런 출력 왜곡과 $(k-1)$번째 층과 k번째 층을 연결하는 가중치 값들의 양자화로 인한 에러는 서로 독립이므로 Δ_1과 Δ_2의 곱에 대한 기대치, $E[\Delta_1\,\Delta_2] = 0$이다. 따라서 Δ_1과 Δ_2의 공분산,

$$Cov[\Delta_1, \Delta_2] = 0 \tag{29}$$

$(k-1)$번째 은닉층의 뉴런 수가 충분히 크다고 가정하면 중심 극한 정리에 의해 Δ_1, Δ_2 그리고 이들의 합인 $\Delta a_i(k)$를 정규 분포를 따르는 랜덤 확률 변수로 취급할 수 있다. 그러면 $\Delta a_i(k)$에 대한 통계적 특성을 다음과 같이 얻을 수 있다.

$$E[\Delta a_i(k)] = \sum_{j=1}^{N_{k-1}} w_{ij}(k-1)\, E[\Delta x_j(k-1)] \tag{30}$$

$$Var[\Delta a_i(k)] = \sum_{j=1}^{N_{k-1}} w_{ij}^2(k-1)\, Var[\Delta x_j(k-1)] + \sigma^2_{\Delta w} \sum_{j=1}^{N_{k-1}} x_j^2(k-1) \tag{31}$$

$$\Delta a_i(k) \sim n(\mu_{\Delta a_i(k)}, \sigma_{\Delta a_i(k)}) \tag{32}$$

위 식에서 $\mu_{\Delta a_i(k)} = E[\Delta a_i(k)]$이고, $\sigma_{\Delta a_i(k)} = \sqrt{Var[\Delta a_i(k)]}$이다.

뉴런의 비선형 활성화 함수를 거친 후의 뉴런 출력 $\hat{t}_i(k)$는 뉴런의 활성화 함수 $T(\cdot)$에 대하여 Taylor series 확장을 적용하여 중심 $a_i(k)$에서 이차 근사까지 취함으로써 다음과 같이 근사할 수 있다.

$$\begin{aligned}\hat{t}_i(k) &= T(a_i(k) + \Delta a_i(k)) \\ &\cong T(a_i(k)) + \Delta a_i(k)\, T'(a_i(k)) + \frac{\Delta a_i^2(k)}{2}\, T''(a_i(k))\end{aligned} \tag{33}$$

그러면 양자화 전의 뉴런 출력 $\hat{t}_i(k)$에 대한 통계적 특성을 다음과 같이 얻을 수 있다.

$$\begin{aligned}E[\hat{t}_i(k)] &= x_i(k) + \frac{1}{2}\, T''(a_i(k))\, Var[\Delta a_i(k)] \\ &\quad + T'(a_i(k))\, E[\Delta a_i(k)]\end{aligned} \tag{34}$$

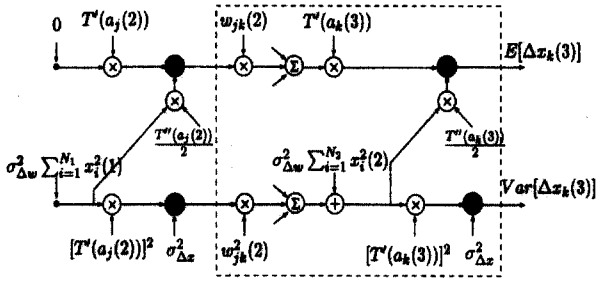
$$Var[\hat{t}_i(k)] = [T'(a_i(k))]^2\, Var[\Delta a_i(k)] \tag{35}$$

마지막으로 k번째 층의 뉴런 출력에 대한 양자화 후의 최종 출력 왜곡 $\Delta x_i(k)$에 대한 기대값과 분산은 다음과 같다.

$$\begin{aligned}E[\Delta x_i(k)] &= \frac{1}{2}\, T''(a_i(k))\, Var[\Delta a_i(k)] \\ &\quad + T'(a_i(k))\, E[\Delta a_i(k)]\end{aligned} \tag{36}$$

$$Var[\Delta x_i(k)] = [T'(a_i(k))]^2\, Var[\Delta a_i(k)] + \sigma^2_{\Delta x} \tag{37}$$

이제까지 분석한 결과를 간략히 정리하여 3층 퍼셉트론 신경망에 대해 도시화하면 (그림 3)와 같다. 검은색으로 표시된 부분이 각 계층의 뉴런 출력단을 의미한다. 은닉층이 2개이상인 다층 퍼셉트론 신경망의 경우, 점선으로 된 박스 안의 내용이 반복되어 나타난다.



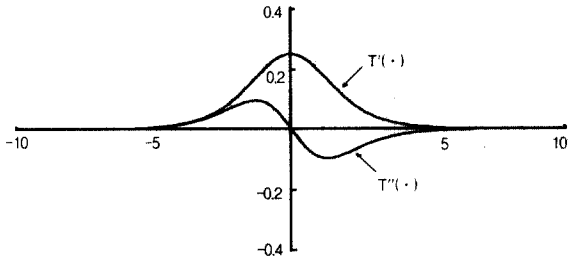
(그림 3) 3층 퍼셉트론 신경망의 양자화 에러 흐름도

3.3 분석에 대한 고찰

이제까지 뉴런 출력과 가중치들의 양자화에 따른 임의의 은닉층 뉴런에서의 출력 왜곡에 대해 자세히 분석하였다. 여기서는 이러한 분석 결과를 토대로 신경망의 출력층 뉴런들의 출력 왜곡에 영향을 미치는 주요 요인들을 점검해 보고, 그 영향을 감소시킬 수 있는 방법에 대하여 고찰해 본다.

(그림 3)의 점선으로 된 박스 안을 보면 이전 층으로부터 전달되는 뉴런들의 입력 제곱들의 합, 가중치 및 가중치 제곱의 합, 뉴런 활성화 함수의 일차 미분치와 이차 미분치가 각 층의 뉴런 출력 왜곡에 대한 평균과 분산에 크게 영향을 미친다는 사실을 알 수 있다. 따라서 이들 세 주요 요인들에 의한 영향을 줄일 수 있는 방법을 강구해야 한다. 첫째, 입력 제곱들의 합을 줄이기 위해서는 각 입력 벡터 요소들이 가능하면 작은 값일수록 좋다. 이에 대한 방안으로는 입력 벡터들을 정규화하는 것이다. 정규화 된 입력 벡터들은 원래의 입력 벡터들이 가지고 있는 정보들을 전혀 손상하지 않

을 뿐 아니라, 학습할 때에도 시간이 덜 걸리는 것으로 알려져 있다. 둘째, 가중치 값과 가중치 값의 제곱을 동시에 작게 하기 위해서는 가중치 값의 절대치를 작게 해야 한다. 가중치의 절대값을 작게 하기 위해서는 학습 도중에 가능하면 가중치 값들이 작은 값을 갖도록 유도하는 방법이 있다. 이는 목표 학습 함수에 가중치 값의 크기에 따른 penalty항을 둠으로써 가능하다. 또 다른 방법으로는 학습 후에 상대적으로 큰 값을 갖는 가중치들을 선택적 증식[11]을 통하여 여러 개의 작은 값을 갖는 가중치들로 분할하는 방법이 있다. 마지막으로 뉴런 활성화 함수의 일차 미분치와 이차 미분치를 동시에 작게 해야 한다. 뉴런 활성화 함수로 시그모이드형태의 로지스틱 함수를 사용할 경우 (그림 4)에서 볼 수 있듯이 이들 값이 동시에 작은 값을 가지는 경우는 로지스틱 함수의 포화 영역밖에 없다. 따라서 학습시 모든 패턴들에 대해 출력층 뉴런들의 출력이 충분히 포화되도록 학습하는 것이 좋다는 것을 알 수 있다.

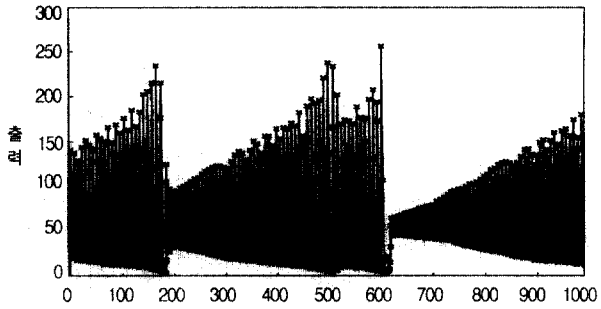


(그림 4) 로지스틱 함수의 일차 및 이차 미분 함수($g=1$)

4. 실 험

본 실험에서는 시계열 예측 방법들의 비교 경연을 위하여 미국의 산타페 연구소(Santa Fe Institute, SFI)에서 제시한 Data Set A인 원적외선 레이저의 출력 데이터[12]를 사용하였다. 이 데이터는 3개의 비선형 상미분 방정식의 결합에 의하여 근사되어질 수 있는 것으로 알려져 있다[13]. (그림 5)은 실험에 사용된 1,000개의 레이저 출력 값을 보여준다. 이 중 처음 800개의 데이터는 학습용 데이터로 사용하였고, 나머지 200개는 테스트용 데이터로 사용하였다. 그리고 현재 시간에 t서 바로 다음의 출력 값 $x(t+1)$을 예측하기 위하여, 신경망의 입력으로 현재 출력 값 $x(t)$를 포함하여 과거 11개의 출력 값들, $x(t-1)$, $x(t-2)$, …, $x(t-11)$을 사용하였다. 실험에는 입력 뉴런 수 12, 은닉층 뉴런 수 5, 그리고 출력층의 뉴런 수가 1인 3층 퍼셉트론 신경망을 사용하였다.

먼저 입력 데이터의 정규화가 가중치 및 뉴런 출력의 양자화에 미치는 영향을 알아보기 위한 실험을 하였다. <표 1>은 정규화하지 않은 원래의 데이터를 사용하여 학습한신경망과 정규화한 데이터를 사용하여 학습한 신경망에 대해 가중치 양자화 정밀도에 따른 출력 왜곡을 보여준다. 이 실험의 결과는 테스트용 데이터를 사용하여 얻어졌고, 초기가중치 값을 서로 다르게 하여 5번의 학습을 통하여 얻어진



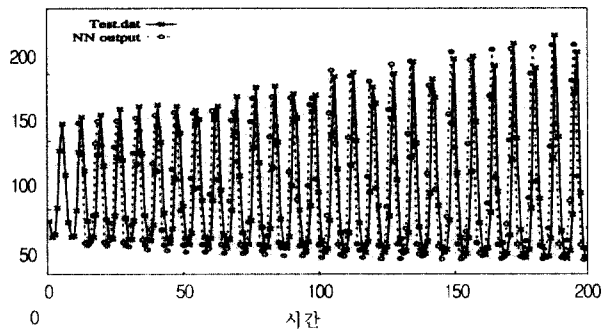
(그림 5) 실험에 사용된 원적외선 레이저의 출력

〈표 1〉 입력 벡터의 정규화 여부에 따른 양자화로 인한 출력 왜곡 ($b_n=8$)

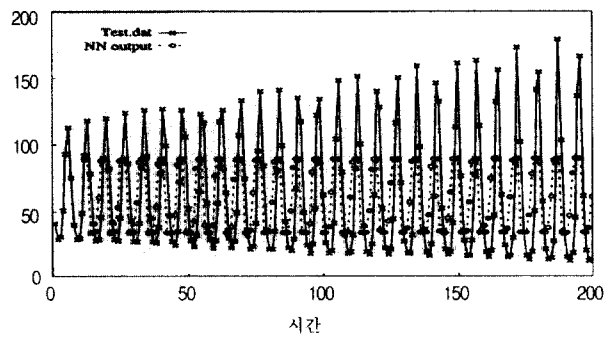
정규화 여부	양자화 정밀도(b_w)					
	16	12	10	8	6	4
비수행	0	0	0.0020	0.0287	0.0831	0.1862
수 행	0	0	0.0003	0.0095	0.0362	0.1251

* 비교를 위하여 정규화 비수행 경우의 출력 왜곡은 정규화하였다.

신경망들에 대한 결과들의 평균을 나타낸다. 한편 뉴런 출력의 양자화 정밀도($4 \le b_n \le 16$)의 변화에 대해서는 전 실험을 통해 출력 왜곡의 어떠한 의미있는 큰 변화는 발견되지 않았다. 따라서 본 논문의 모든 실험에서는 $b_n = 8$ 로 고정시키고 수행하였다. 입력 데이터에 대해 정규화를 수행한 경우에 비해 정규화를 수행하지 않은 경우에서 양자화 정밀도가 낮아짐에 따라 출력 왜곡의 정도가 더 심해지는 것을 알 수 있다. (그림 6)은 정규화하지 않은 원래의 데이



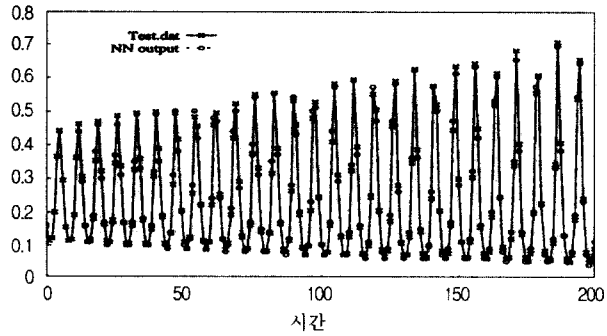
(a) $b_w=8$, $b_n=8$



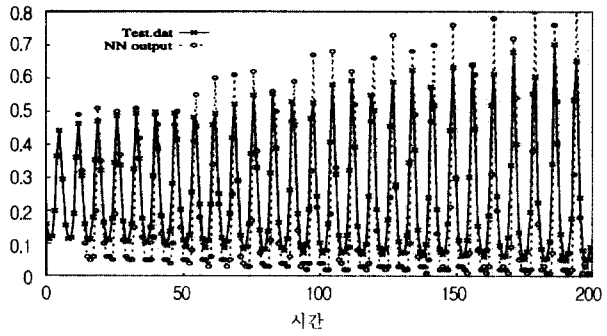
(b) $b_w=4$, $b_n=8$

(그림 6) 원래 데이터를 사용해 학습한 신경망의 출력

터를 사용하여 학습한 신경망의 가중치 양자화를 위한 비트 수가 각각 8과 4일 때 양자화 후의 출력 예를 보여준다. 그리고 (그림 7)은 입력 데이타들을 [0, 1] 사이로 정규화한 데이터를 사용하여 학습한 신경망의 가중치 양자화 비트 수가 각각 8과 4일 때 양자화 후의 출력 예를 보여준다.



(a) $b_w=8,\ b_n=8$



(b) $b_w=4,\ b_n=8$

(그림 7) 정규화된 데이터를 사용하여 학습한 신경망의 출력

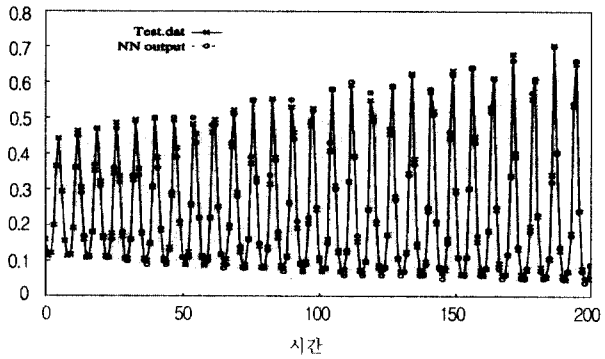
〈표 2〉 W_{max}의 변화에 따른 증식된 신경망의 뉴런 수

	W_{max}				
	무제한	8	4	2	1
입력층 뉴런 수	12	13	15	21	32
은닉층 뉴런 수	5	6	8	12	19
평균 출력 왜곡	0.1246	0.1037	0.0874	0.0883	0.0946

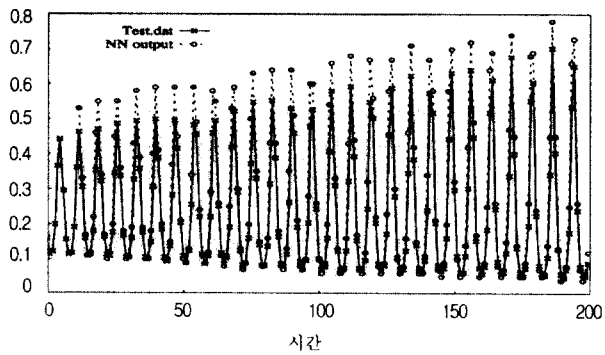
다음으로 작은 크기의 가중치 값들을 갖는 신경망이 양자화로 인한 출력 왜곡에 얼마나 덜 민감한 지를 실험하였다. 먼저 학습용 데이터와 테스트용 데이터에 대해 모두 [0, 1]이내로 정규화하였다. 그리고 weight decay방법을 사용하여 신경망을 충분히 학습시켰다. 그 다음 학습된 신경망을 W_{max}의 다양한 값에 대해 선택적 증식[11]을 수행하였다. 선택적 증식의 경우 가중치의 최대 허용 크기 W_{max}의 적절한 값을 선택하는 것이 중요하다. <표 2>는 W_{max} 값의 변화에 따른 선택적 증식 후의 입력층 및 은닉층 뉴런의 수 그리고 출력층 뉴런의 평균 출력 왜곡을 보여준다. 이 실험의 결과는 테스트용 데이터를 사용하여 얻어졌고, 초기 가중치 값을 서로 다르게 하여 학습을 통하여 얻어진 5개의 신경망들에 대한 결과들의 평균을 나타낸다. 그리고 <표 2>에서 평균 출력 왜곡은 $b_w=4$에 대하여 수행된 결과이다. 이 실험 결과로부터 적절한 W_{max}의 값으로 4를 선택하였다. <표 3>은 $W_{max}=4$일 때 선택적 증식에 의해 얻어진 신경망들의 양자화 정밀도의 변화에 따른 평균 출력 왜곡을 보여준다. 앞의 <표 1>과 비교하여 양자화로 인한 출력 왜곡이 꽤 줄어들었음을 볼 수 있다. 그리고 (그림 8)는 선택적 증식을 수행하여 얻어진 신경망의 가중치 양자화 비트 수가 각각 8과 4일 때 양자화 후의 출력 예를 보여준다. 역시 앞의 (그림 7)의 경우와 비교하여 훨씬 안정된 출력을 보임을 알 수 있다.

〈표 3〉 양자화 정밀도에 따른 출력 왜곡($W_{max}=4$)

	양자화 정밀도(비트 수, b_w)					
	16	12	10	8	6	4
평균 출력 왜곡	0	0	0	0.0048	0.0127	0.0874



(a) $b_w=8,\ b_n=8$



(b) $b_w=4,\ b_n=8$

(그림 8) 제안된 방법에 따른 신경망의 출력($W_{max}=4$)

5. 결 론

인공 신경망을 실제 응용 분야에 다양하게 적용하기 위해서는 이를 하드웨어로 구현하는 것이 필요하다. 하드웨어로 구현하는 방법에는 현재 하이브리드 VLSI 신경망 칩으로 구현하는 것이 가장 유망하다. 이미 학습된 신경망을 디지털

또는 하이브리드 신경망 칩을 사용하여 구현하는 경우 뉴런 출력과 가중치 값의 양자화 과정이 따라야 한다. 이것은 신경망의 출력층 뉴런 출력에서 이미 학습된 출력에 비해 왜곡을 야기한다. 본 논문에서는 이러한 신경망의 출력 왜곡을 자세하게 분석하였다. 그리고 분석 결과를 토대로 신경망의 출력 왜곡을 줄일 수 있는 한 방법으로 입력 벡터의 정규화와 학습을 통하여 신경망이 작은 가중치 값을 갖는 해를 선택할 것을 제시하였고 학습 후 선택적 증식 변환을 수행하는 것이 필요하다는 것을 알았다. 이러한 결과는 Xie와 Jabri가 주장한 신경망 출력 왜곡에 대해 가중치의 영향이 크다는 사실과 일치한다. 또한 선택적 증식 방법을 통하여 Stevenson과 Widrow가 주장한 뉴런당 가중치 수와 계층당 뉴런의 수가 신경망 출력 왜곡과는 무관하다는 사실과도 일치하는 것이다. 실험 결과는 이러한 방법들을 통해 실제로 뉴런 출력 및 가중치 값의 양자화로 인한 출력층 뉴런의 출력 왜곡을 상당히 줄일 수 있음을 명확히 보여 주었다.

참 고 문 헌

[1] B. Widrow, D. E. Rumelhart, and M. A. Lehr, "Neural Networks : Applications in industry, business and science," Communications of the ACM, Vol.37, No.3, pp.93-105, Mar., 1994.

[2] 포항공대, "신경망 칩 응용 기반 기술 연구", 한국통신(KT-93-45) 장기 기초 연구 과제 보고서, Dec., 1993.

[3] E. Sackinger, B. E. Boser, J. Bromley, Y. LeCun, and L. D. Jackel, "Application of the ANNA Neural Network Chip to High Speed Character Recognition," IEEE Transactions on Neural Networks, Vol.3, No.3, pp.498-505, 1992.

[4] Maryhelen Stevenson, Rodney Winter and Bernard Widrow, "Sensitivity of Feedforward Neural Networks to Weight Errors," IEEE Trans. on Neural Networks, Vol.1, pp.71-80, 1990.

[5] Yun Xie and Marwan A. Jabri, "Analysis of the Effects of Quantization in Multilayer Neural Networks Using a Statistical Model," IEEE Trans. on Neural Networks, Vol.3, pp.334-338, 1992.

[6] Stephen W. Pichè, "The Selection of Weight Accuracies for Madaline," IEEE Trans. on Neural Networks, Vol.6, pp. 432-445, 1995.

[7] Jin-Young Choi and Chong-Ho Choi, "Sensitivity Analysis of Multilayer Perceptron with Differentiable Activation Functions," IEEE Trans. on Neural Networks, Vol.3, pp. 101-107, 1992.

[8] D. Lovell, P. Bartlett and T. Downs, "Error and Variance Bounds on Sigmoidal Neurons with Weight and Input Errors," Electronics Letters, Vol.28, pp.760-762, 1992.

[9] Jordan L. Holt and Jenq-Neng Hwang, "Finite Precision Error Analysis of Neural Network Hardware Implementations," IEEE Trans. on Computers, Vol.42, pp.281-290, 1993.

[10] A. Papoulis, Probability, Random Variables, and Stochastic Processes, McGraw-Hill, 1987.

[11] Oh-Jun Kwon and Sung-Yang Bang, "Design of a Fault Tolerant Neural Network with a Desired Level of Robustness," Electronics Letters, Vol.33, No.12, pp.1055-1057, 1997.

[12] Neil A. Gershenfeld and Andreas S. Weigend, "The Future of Time Series : Learning and Understanding," In Time Series Prediction : Forecasting the Future and Understanding the Past, Addison Wesley, 1993.

[13] Udo übner, Carl-Otto Weiss, Neal Broadus Abraham, and Dingyuan Tang, "Lorenz-Like Chaos in NH_3-FIR Lasers(Data Set A)," In Time Series Prediction : Forecasting the Future and Understanding the Past, Addison Wesley, 1993.

권 오 준

e-mail : ojkwon@dongeui.ac.kr

1986년 경북대학교 전자공학과 졸업 (공학사)

1992년 충남대학교 대학원 전산학과 (이학석사)

1998년 포항공과대학교 대학원 전자계산학과(공학박사)

1986년~2000년 한국전자통신연구원 선임연구원

2002년~현재 동의대학교 컴퓨터·영상공학부 조교수

관심분야 : 지능정보 처리, 신경망 응용, 패턴 인식, 컴퓨터 통신망 및 정보 보호, 차세대 인터넷

김 성 우

e-mail : libero@dongeui.ac.kr

1991년 한국과학기술원 전기 및 전자공학과 (공학사)

1993년 한국과학기술원 전기 및 전자공학과 (공학석사)

1999년 한국과학기술원 전기 및 전자공학과 (공학박사)

1999년~2001년 한국전자통신연구원 컴퓨터소프트웨어기술연구소 선임연구원

2002년~현재 동의대학교 컴퓨터·영상공학부 전임강사

관심분야 : 실시간 운영체제, 그래픽 윈도우 시스템, 내고장성 시스템

이 종 민

e-mail : jongmin@dongeui.ac.kr

1992년 경북대학교 컴퓨터공학과(공학사)

1994년 한국과학기술원 전산학과(공학석사)

2000년 한국과학기술원 전산학과(공학박사)

1997년~2002년 삼성전자 책임연구원

2002년~현재 동의대학교 컴퓨터·영상공학부 전임강사

관심분야 : 모바일 컴퓨팅, 인터넷 프로토콜, 병렬처리