(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2019년10월15일
(11) 등록번호 10-2032146
(24) 등록일자 2019년10월08일

(51) 국제특허분류(Int. Cl.)
G06N 3/04 (2006.01) *G06N 3/08* (2006.01)
(52) CPC특허분류
G06N 3/04 (2013.01)
G06N 3/08 (2013.01)
(21) 출원번호 10-2018-0042350
(22) 출원일자 2018년04월11일
심사청구일자 2018년04월11일
(56) 선행기술조사문헌
Xiaojie Jin 외 5명. Deep Learning with S-shaped Rectified Linear Activation Units. 2015.12.22.
KR1020170136357 A
KR1020180027972 A
KR1020160063965 A

(73) 특허권자
경희대학교 산학협력단
경기도 용인시 기흥구 덕영대로 1732 (서천동, 경희대학교 국제캠퍼스내)
(72) 발명자
배성호
대전광역시 서구 신갈마로113번길 40, 106동 701호
이승룡
경기도 성남시 분당구 구미로144번길 25, 102동 102호(구미동, 삼환빌라)
(74) 대리인
김홍석

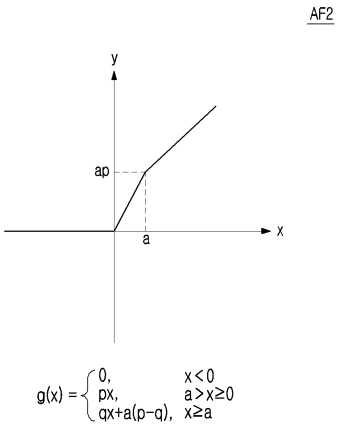
전체 청구항 수 : 총 11 항 심사관 : 최정권

(54) 발명의 명칭 **소자 결점을 보완하기 위한 구간 선형 정류 유닛을 사용하는 인공신경망 시스템**

(57) 요 약

본 발명의 인공신경망 시스템은 입력 데이터를 수신하는 입력층, 상기 입력 데이터에 대응하는 출력 데이터를 출력하는 출력층, 그리고 상기 입력층과 상기 출력층 사이에 위치하는 복수의 은닉층들을 포함하며, 상기 은닉층들 사이는 연결 가중치들을 통해 연결되고, 상기 연결 가중치들은 상기 출력 데이터와 상기 입력 데이터에 대응하는 예측 데이터의 차이에 기초하여 갱신되는 은닉층 그룹을 포함하되, 상기 은닉층들 각각은 활성화 함수를 포함하는 은닉 노드를 포함하고, 상기 활성화 함수는 입력값이 음수인 경우 0을 출력하고, 입력값이 양수인 경우 입력값의 범위에 따라 기울기가 서로 다른 선형 함수들에 기초하여 출력값을 결정하고, 상기 선형 함수들은 원점에서 멀어질수록 기울기가 감소할 수 있다.

대 표 도 - 도5



이 발명을 지원한 국가연구개발사업

과제고유번호	10085646
부처명	산업통상자원부
연구관리전문기관	산업기술평가관리원(KEIT)
연구사업명	산업기술혁신사업
연구과제명	멤리스터 어레이 결점 극복을 위한 뉴로모픽 시스템
기 여 율	1/1
주관기관	경희대학교산학협력단
연구기간	2017.12.01 ~ 2018.12.31

명 세 서

청구범위

청구항 1

입력 데이터를 수신하는 입력층;

상기 입력 데이터에 대응하는 출력 데이터를 출력하는 출력층; 그리고

상기 입력층과 상기 출력층 사이에 위치하는 복수의 은닉층들을 포함하며, 상기 은닉층들 사이는 연결 가중치들을 통해 연결되고, 상기 연결 가중치들은 상기 출력 데이터와 상기 입력 데이터에 대응하는 예측 데이터의 차이에 기초하여 갱신되는 은닉층 그룹을 포함하되,

상기 은닉층들 각각은 활성화 함수를 포함하는 은닉 노드를 포함하고,

상기 활성화 함수는 입력값이 음수인 경우 0을 출력하고, 입력값이 양수인 경우 입력값의 범위에 따라 기울기가 서로 다른 선형 함수들에 기초하여 출력값을 결정하고,

상기 선형 함수들은 원점에서 멀어질수록 기울기가 감소하는 인공신경망 시스템.

청구항 2

제1항에 있어서,

상기 활성화 함수는 아래의 수학식에 따라 정의되고,

입력값은 원점 및 제1 파라미터(a)에 의해 구분되고,

상기 활성화 함수는 원점과 상기 제1 파라미터(a) 사이의 제1 구간에서 제1 기울기(p)를 가지고, 상기 제1 파라미터(a)보다 큰 제2 구간에서 제2 기울기(q)를 가지고,

상기 제2 기울기(q)는 상기 제1 기울기(p)보다 작은 것을 특징으로 하는 인공신경망 시스템.

$$g(x)=\begin{cases}0, & x<0\\ px, & a>x\geq 0\\ qx+a(p-q), & x\geq a\end{cases}$$

청구항 3

제2항에 있어서,

상기 제1 파라미터(a)는 원점으로부터 순차적으로 증가하면서 상기 출력 데이터와 상기 예측 데이터 사이의 차이가 최소로 되는 값으로 결정되는 인공신경망 시스템.

청구항 4

제3항에 있어서,

상기 제1 기울기(p) 및 상기 제2 기울기(q)는 오류역전파 학습 시 경사 하강법을 이용하여 결정되는 인공신경망 시스템.

청구항 5

제1항에 있어서,

상기 활성화 함수는 아래의 수학식에 따라 정의되고,

입력값은 원점, 제1 파라미터(a) 및 제2 파라미터(b)에 의해 구분되고,

상기 활성화 함수는 원점과 상기 제1 파라미터(a) 사이의 제1 구간에서 제1 기울기(p)를 가지고, 상기 제1 파라미터(a)와 상기 제2 파라미터(b) 사이의 제2 구간에서 제2 기울기(q)를 가지고, 상기 제2 파라미터(b)보다 큰 제3 구간에서 제3 기울기(r)를 가지고,

상기 제3 기울기(r)는 상기 제2 기울기(q)보다 작고, 상기 제2 기울기(q)는 상기 제1 기울기(p)보다 작은 것을 특징으로 하는 인공신경망 시스템.

$$h(x) = \begin{cases} 0, & x < 0 \\ px, & a > x \geq 0 \\ qx + a(p-q), & b > x \geq a \\ rx + a(p-q) + b(q-r), & x \geq b \end{cases}$$

청구항 6

제5항에 있어서,

상기 제1 파라미터(a)는 원점으로부터 순차적으로 증가하면서 상기 출력 데이터와 상기 예측 데이터 사이의 차이가 최소로 되는 값으로 결정되는 인공신경망 시스템.

청구항 7

제6항에 있어서,

상기 제2 파라미터(b)는 상기 제1 파라미터(a)로부터 순차적으로 증가하면서 상기 출력 데이터와 상기 예측 데이터 사이의 차이가 최소로 되는 값으로 결정되는 인공신경망 시스템.

청구항 8

제7항에 있어서,

상기 제1 기울기(p), 상기 제2 기울기(q) 및 상기 제3 기울기(r)는 오류역전파 학습 시 경사 하강법을 이용하여 결정되는 인공신경망 시스템.

청구항 9

입력 데이터를 수신하는 적어도 하나의 입력 노드;

상기 입력 데이터에 대응하는 출력 데이터를 출력하는 적어도 하나의 출력 노드; 그리고

상기 입력 노드와 상기 출력 노드 사이에 위치하는 복수의 은닉 노드들을 포함하되,

상기 은닉 노드들은 서로 연결 가중치들을 통해 연결되고, 상기 연결 가중치들은 상기 출력 데이터와 상기 입력 데이터에 대응하는 예측 데이터의 차이에 기초하여 갱신되며,

상기 은닉 노드들 각각은 노드의 출력 범위를 제어하는 활성화 함수를 포함하고,

상기 활성화 함수는 복수의 파라미터들에 의해 입력값의 범위를 분할하고, 분할된 구간들에서 서로 다른 기울기를 가지는 선형 함수들에 기초하여 출력값을 결정하고,

상기 선형 함수들은 원점에서 멀어질수록 기울기가 감소하는 인공신경망 시스템.

청구항 10

제9항에 있어서,

상기 파라미터들은 상기 출력 데이터와 상기 예측 데이터 사이의 차이가 최소로 되는 값으로 결정되는 인공신경망 시스템.

청구항 11

제10항에 있어서,

상기 선형 함수들 각각의 기울기는 오류역전파 학습 시 경사 하강법을 이용하여 결정되는 인공신경망 시스템.

발명의 설명

기 술 분 야

[0001] 본 발명은 인공신경망 시스템에 관한 것으로, 좀 더 구체적으로는 소자 결점을 보완하기 위한 구간 선형 정류 유닛을 사용하는 인공신경망 시스템에 관한 것이다.

배 경 기 술

[0002] 인공신경망은 논리식을 기본으로 하고 프로그램에 의해 동작하는 폰 노이만(von Neumman)형 컴퓨터 구조와 달리 인간의 뇌를 모방하여 만든 연산 방법이다. 인간의 뇌가 컴퓨터에 대해서 보이는 차이점은 크게 두 가지로 요약할 수 있다. 첫 번째는 대규모 병렬 처리 및 분산 처리 형태(Massively distributed and parallel)로 연산을 수행한다는 것이고, 두 번째는 문제를 위한 선행적 프로그램 개발 대신에 학습을 통해서 문제를 해결한다는 것이다. 이러한 인간의 뇌가 가지는 특성을 흉내 내는 연산 방법이 바로 인공신경망이다.

[0003] 하드웨어 인공신경망의 개발은 폰 노이만형 컴퓨터에 의한 디지털 컴퓨터의 개발과 그 시기를 비슷하게 갖고 있다. 다시 말해서 인공신경망 컴퓨터는 기존의 폰 노이만형 컴퓨터의 대치를 위해 개발된 것이 아니라, 인간의 뇌 구조를 모방하려는 노력에서 나온 폰 노이만 형 컴퓨터와는 또 다른 종류의 하나인 것이다. 이러한 인공신경망 컴퓨터의 첫 번째 하드웨어 구현은 F. Rosenblatt에 의해 1958년에 개발된 MARK I Perceptron이었다. 이후 40년 동안 인공신경망은 발전과 쇠퇴를 반복해왔으나, 디지털 컴퓨터가 IC의 출현에 힘입어 급속도로 발전한 반면 인공신경망 컴퓨터는 일반화된 문제에 적용될 만큼 발전하지 못했었다.

[0004] 이러한 초창기의 인공신경망은 대부분 아날로그 기술로 만들어졌다. 일반적으로 아날로그 인공신경망은 논리적인 인공신경망을 직관적으로 만들 수 있는 장점이 있으며, 가중치 적용에 대해서 곱셈기를 사용하지 않음으로써 적은 면적을 차지하면서도 매우 빠른 병렬 구조를 획득할 수 있다. 반면에, 범용성이 크게 제한되고 학습된 가중치 값의 저장이 어려우며, 전기적 잡음과 온도에 민감하여 오차율이 크고, 기존의 디지털 컴퓨터와의 인터페이스가 어려운 단점들이 있어서 상업적 제품으로는 성공하지 못하였다. 이후 90년대 중반까지 이러한 인공신경망 칩 개발은 다소 침체된 양상을 보였었으나, 90년대 후반에 들어서 디지털 반도체 기술의 비약적인 발전에 힘입어 인공신경망 칩에 대한 연구가 다시 활성화되기 시작했다. ASIC 기술이 발전하고 FPGA가 등장함에 따라 빠르고 집적도 높은 디지털 VLSI의 구현이 가능해졌고, 이를 이용한 인공신경망 칩도 많이 개발되어 상업용 인공신경망 칩도 활발히 개발되었다. 또한, 고속 DSP의 등장은 병렬 DSP를 이용한 인공신경망의 구현을 가능하게 하였다. 아날로그 인공신경망 칩에 비해 면적을 많이 차지하고 느린 속도를 가진다는 단점이 있으나, 가중치 값의 저장이 용이하고 기존의 디지털 프로세서 기술을 사용함으로써 쉽게 제작할 수 있으며, 아날로그 인공신경망에 비해 비교적 자유로운 구조를 가질 수 있는 장점이 있다.

[0005] 디지털 인공신경망을 구현하는 방법은 크게 두 가지로 나눌 수 있다. 첫 번째는 DSP 혹은 CPU를 이용한 마이크로프로세서 기반의 인공신경망이며, 두 번째는 ASIC 형태로써 직접적으로 시냅스, 뉴런 등의 회로를 구성하여 인공신경망을 구현하는 방법이다. 디지털 프로세서 기반의 인공신경망은 구현되는 시냅스의 개수에 비해 면적이 적고, 기존의 프로세서를 그대로 사용하기 때문에 설계가 용이하며, 프로그램의 변경만으로 다양한 형태의 인공신경망을 구현할 수 있다는 장점이 있다. 하지만, ASIC으로 구현되는 인공신경망에 비해서 속도가 느리고, 단일 칩으로 구성하기 어려우며, 신경망의 특징인 병렬처리, 분산처리의 효율이 떨어지게 된다. ASIC 기술을 이용하여 인공신경망을 구현하는 경우는 사용하고자 하는 목적에 따라서 다양한 형태의 인공신경망을 구현하는 것이 가능하며 이론적 모델에 가까운 형태로 신경망을 구현할 수 있다는 장점이 있지만, 한번 구현된 신경망의 형태를 변경하기가 어렵고, 신경망의 확장이 어려운 단점이 있다.

[0006] 일반적인 완전연결(Fully Connected) 형태를 갖는 신경망 구조에서 뉴런의 증가에 대해 시냅스의 개수는 뉴런 수의 제곱의 형태로 증가하게 되고, 이에 비례하여 곱셈기의 양이 증가하게 된다. 이러한 곱셈기의 증가에 따른 면적 소모 및 속도 저하는 곱셈기가 사용되지 않는 아날로그 인공신경망에서보다 디지털 인공신경망에서 더 큰 단점으로 나타나게 되며, 이를 극복하기 위한 방법으로 곱셈기가 없는 디지털 신경망, 혹은 곱셈기의 크기를 줄이기 위한 시리얼 데이터 방식의 디지털 신경망과 같이 구조를 그대로 두고 곱셈기의 크기를 줄이는 연구들이 이루어졌다. 그리고 다른 방법으로는 SIMD 형태와 같이 버스 구조를 통해서 곱셈기를 재사용하는 디지털 신경망과 사슬 배열 구조(Systolic Array Architecture)를 채용하여 곱셈기의 수를 선형적으로 늘이는 인공신경망, 셀룰러 인공신경망(Cellular Neural Network)처럼 인접 연결(Neighborhood Connected)을 사용하는 신경망과 같이

기존의 완전 연결 구조를 물리적으로 구현하지 않고 모듈러 아키텍처 형태로 구현함으로써 뉴런 및 시냅스의 증가에 따른 곱셈기의 증가 문제를 해결하는 연구가 이루어졌다.

[0007] 현재 대부분의 인공신경망은 CMOS(Complementary Metal-Oxide Semiconductor)를 활용하여 구현되고 있다. 하지만, 고집적, 저전력의 딥러닝 구조를 상용 하드웨어로 구현하기 위해서는 CMOS를 대체할 새로운 소자(예를 들어, ReRAM)의 사용이 필요할 것으로 예상되고 있다. 다만, 고집적, 저전력 소자들은 채널 간섭 현상 및 불안정한 다층 구조로 인해 읽기 동작 시 왜곡된 값을 인출할 수 있다. 따라서, 이러한 소자의 오류를 보완하기 위한 방법이 필요하다.

선행기술문헌

특허문헌

[0008] (특허문헌 0001) 대한민국 공개특허 제10-2017-0140225호

발명의 내용

해결하려는 과제

[0009] 본 발명의 목적은 소자 오류로 인한 왜곡을 보완하기 위해 변형된 활성화 함수를 사용하는 인공신경망 시스템을 제공하는 데 있다.

과제의 해결 수단

[0010] 본 발명의 실시 예에 따른 인공신경망 시스템은 입력 데이터를 수신하는 입력층, 상기 입력 데이터에 대응하는 출력 데이터를 출력하는 출력층, 그리고 상기 입력층과 상기 출력층 사이에 위치하는 복수의 은닉층들을 포함하며, 상기 은닉층들 사이는 연결 가중치들을 통해 연결되고, 상기 연결 가중치들은 상기 출력 데이터와 상기 입력 데이터에 대응하는 예측 데이터의 차이에 기초하여 갱신되는 은닉층 그룹을 포함하되, 상기 은닉층들 각각은 활성화 함수를 포함하는 은닉 노드를 포함하고, 상기 활성화 함수는 입력값이 음수인 경우 0을 출력하고, 입력값이 양수인 경우 입력값의 범위에 따라 기울기가 서로 다른 선형 함수들에 기초하여 출력값을 결정하고, 상기 선형 함수들은 원점에서 멀어질수록 기울기가 감소할 수 있다.

[0011] 실시 예로서, 상기 활성화 함수는 아래의 수학식에 따라 정의되고, 입력값은 원점 및 제1 파라미터(a)에 의해 구분되고, 상기 활성화 함수는 원점과 상기 제1 파라미터(a) 사이의 제1 구간에서 제1 기울기(p)를 가지고, 상기 제1 파라미터(a)보다 큰 제2 구간에서 제2 기울기(q)를 가지고, 상기 제2 기울기(q)는 상기 제1 기울기(p)보다 작은 것을 특징으로 할 수 있다.

$$g(x)=\begin{cases}0, & x<0\\ px, & a>x\geq 0\\ qx+a(p-q), & x\geq a\end{cases}$$

[0012]

[0013] 실시 예로서, 상기 제1 파라미터(a)는 원점으로부터 순차적으로 증가하면서 상기 출력 데이터와 상기 예측 데이터 사이의 차이가 최소로 되는 값으로 결정될 수 있다.

[0014] 실시 예로서, 상기 제1 기울기(p) 및 상기 제2 기울기(q)는 오류역전파 학습 시 경사 하강법을 이용하여 결정될 수 있다.

[0015] 실시 예로서, 상기 활성화 함수는 아래의 수학식에 따라 정의되고, 입력값은 원점, 제1 파라미터(a) 및 제2 파라미터(b)에 의해 구분되고, 상기 활성화 함수는 원점과 상기 제1 파라미터(a) 사이의 제1 구간에서 제1 기울기(p)를 가지고, 상기 제1 파라미터(a)와 상기 제2 파라미터(b) 사이의 제2 구간에서 제2 기울기(q)를 가지고, 상기 제2 파라미터(b)보다 큰 제3 구간에서 제3 기울기(r)를 가지고, 상기 제3 기울기(r)는 상기 제2 기울기(q)보다 작고, 상기 제2 기울기(q)는 상기 제1 기울기(p)보다 작은 것을 특징으로 할 수 있다.

$$h(x) = \begin{cases} 0, & x<0 \\ px, & a>x\geq 0 \\ qx+a(p-q), & b>x\geq a \\ rx+a(p-q)+b(q-r), & x\geq b \end{cases}$$

[0016]

[0017] 실시 예로서, 상기 제1 파라미터(a)는 원점으로부터 순차적으로 증가하면서 상기 출력 데이터와 상기 예측 데이터 사이의 차이가 최소로 되는 값으로 결정될 수 있다.

[0018] 실시 예로서, 상기 제2 파라미터(b)는 상기 제1 파라미터(a)로부터 순차적으로 증가하면서 상기 출력 데이터와 상기 예측 데이터 사이의 차이가 최소로 되는 값으로 결정될 수 있다.

[0019] 실시 예로서, 상기 제1 기울기(p), 상기 제2 기울기(q) 및 상기 제3 기울기(r)는 오류역전파 학습 시 경사 하강법을 이용하여 결정될 수 있다.

[0020] 본 발명의 실시 예에 따른 인공신경망 시스템은 입력 데이터를 수신하는 적어도 하나의 입력 노드, 상기 입력 데이터에 대응하는 출력 데이터를 출력하는 적어도 하나의 출력 노드, 그리고 상기 입력 노드와 상기 출력 노드 사이에 위치하는 복수의 은닉 노드들을 포함하되, 상기 은닉 노드들은 서로 연결 가중치들을 통해 연결되고, 상기 연결 가중치들은 상기 출력 데이터와 상기 입력 데이터에 대응하는 예측 데이터의 차이에 기초하여 갱신되며, 상기 은닉 노드들 각각은 노드의 출력 범위를 제어하는 활성화 함수를 포함하고, 상기 활성화 함수는 복수의 파라미터들에 의해 입력값의 범위를 분할하고, 분할된 구간들에서 서로 다른 기울기를 가지는 선형 함수들에 기초하여 출력값을 결정하고, 상기 선형 함수들은 원점에서 멀어질수록 기울기가 감소할 수 있다.

[0021] 실시 예로서, 상기 파라미터들은 상기 출력 데이터와 상기 예측 데이터 사이의 차이가 최소로 되는 값으로 결정될 수 있다.

[0022] 실시 예로서, 상기 선형 함수들 각각의 기울기는 오류역전파 학습 시 경사 하강법을 이용하여 결정될 수 있다.

발명의 효과

[0023] 본 발명의 실시 예에 따르면, 소자 오류로 인한 왜곡을 보완하기 위해 변형된 활성화 함수를 사용하는 인공신경망 시스템을 제공할 수 있다.

도면의 간단한 설명

[0024] 도 1은 본 발명의 실시 예에 따른 인공신경망 시스템을 보여주는 블록도이다.

도 2는 도 1의 인공신경망 시스템을 자세히 보여주는 도면이다.

도 3은 도 2의 하나의 은닉 노드를 보여주는 도면이다.

도 4는 일반적인 활성화 함수인 ReLU(Rectified Linear Unit)를 보여주는 도면이다.

도 5는 본 발명의 실시 예에 따른 활성화 함수를 보여주는 도면이다.

도 6은 본 발명의 다른 실시 예에 따른 활성화 함수를 보여주는 도면이다.

발명을 실시하기 위한 구체적인 내용

[0025] 앞의 일반적인 설명 및 다음의 상세한 설명 모두 예시적이라는 것이 이해되어야 하며, 청구된 발명의 부가적인 설명이 제공되는 것으로 여겨져야 한다. 참조부호들이 본 발명의 바람직한 실시 예들에 상세히 표시되어 있으며, 그것의 예들이 참조 도면들에 표시되어 있다. 가능한 어떤 경우에도, 동일한 참조 번호들이 동일한 또는 유사한 부분을 참조하기 위해서 설명 및 도면들에 사용된다.

[0026] 제1 또는 제2 등의 용어를 다양한 구성요소들을 설명하는데 사용될 수 있지만, 상기 구성요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만, 예를 들어 본 발명의 개념에 따른 권리 범위로부터 이탈되지 않은 채, 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제 2 구성요소는 제 1 구성요소로도 명명될 수 있다.

[0027] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성요소에

직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성요소가 다른 구성요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는, 중간에 다른 구성요소가 존재하지 않는 것으로 이해되어야 할 것이다. 구성요소들 간의 관계를 설명하는 표현들, 예를 들어 "~사이에"와 "바로~사이에" 또는 "~에 직접 이웃하는" 등도 마찬가지로 해석되어야 한다.

[0028] 본 명세서에서 사용한 용어는 단지 특정한 실시 예들을 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 설시된 특징, 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함으로 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0029] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 갖는 것으로 해석되어야 하며, 본 명세서에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0030] 이하, 실시 예들을 첨부된 도면을 참조하여 상세하게 설명한다. 그러나 특허출원의 범위가 이러한 실시 예들에 의해 제한되거나 한정되는 것은 아니다. 각 도면에 제시된 동일한 참조 부호는 동일한 부재를 나타낸다.

[0031] 도 1은 본 발명의 실시 예에 따른 인공신경망 시스템을 보여주는 블록도이다. 도 2는 도 1의 인공신경망 시스템을 자세히 보여주는 도면이다. 도 1 및 도 2를 참조하면, 인공신경망 시스템(100)은 입력층(110), 복수의 은닉층들(120) 및 출력층(130)은 포함할 수 있다. 인공신경망 시스템(100)은 연결선으로 연결된 복수의 인공 뉴런들을 이용하여 생물학적인 시스템의 계산 능력을 모방하는 소프트웨어나 하드웨어로 구현될 수 있다.

[0032] 인공신경망 시스템(100)은 입력(IN)에 대응하는 출력(OUT)을 제공할 수 있다. 예를 들면, 입력(IN)은 음성 데이터, 영상 데이터, 생체 데이터 및 필적 데이터 등을 포함할 수 있다. 인공신경망 시스템(100)은 입력된 음성 데이터가 어떠한 문장(또는 단어)인지를 인식하거나, 동영상에 나타난 인물이 누구인지를 인식할 수 있다. 또한, 인공신경망 시스템(100)은 심전도(ECG; electrocardiogram), 뇌전도(EEG; electroencephalogram) 등의 생체 신호를 분석하여 사용자의 신체 상태를 인식하거나, 또는 사용자의 동작을 분석하여 입력된 필적을 인식할 수 있다. 한편, 인공신경망 시스템(100)은 DNA(deoxyribonucleic acid) 염기 서열 결정 장치(DNA sequencing device)에서 관찰 신호로부터 올바른 DNA 염기 서열을 추정하는데 이용될 수 있다. 즉, 출력(OUT)은 입력된 데이터에 대한 인식 결과일 수 있다.

[0033] 인공신경망 시스템(100)은 복수의 층들을 포함할 수 있다. 예를 들면, 입력층(110)은 학습을 수행하기 위한 입력(IN)을 수신하여 복수의 은닉층들(120)로 전달할 수 있다. 복수의 은닉층들(120)은 제1 은닉층(121) 내지 제n 은닉층(12n)을 포함할 수 있다. 복수의 은닉층들(120)은 입력층(110)을 통해 전달된 학습 데이터를 예측하기 쉬운 값으로 변화시킬 수 있다. 각 은닉층(121~12n)은 이전 은닉층에서 수신된 입력 값을 특정한 기준에 기초하여 다음 은닉층으로 전달하고, 전달 시 은닉층마다 서로 다른 연결 가중치를 적용하여 전달할 수 있다. 출력층(130)은 복수의 은닉층들(120)로부터 수신한 신호에 기초하여 출력(OUT)을 생성할 수 있다. 입력층(120), 각 은닉층(121~12n) 및 출력층(130)에서, 각 층 내의 직접적인 연결은 존재하지 않고, 각 층 사이의 연결만 존재할 수 있다.

[0034] 즉, 인공신경망 시스템(100)은 생물학적인 뉴런의 기능을 단순화시킨 인공 뉴런들이 이용되고, 인공 뉴런들은 연결 가중치를 가지는 연결선을 통해 상호 연결될 수 있다. 연결 가중치는 연결선이 갖는 특정한 값으로서 연결 강도라고도 표현할 수 있다. 인공 뉴런은 노드로 지칭될 수 있다.

[0035] 입력층(110), 각 은닉층(121~12n) 및 출력층(130)은 복수의 노드들을 포함할 수 있다. 예를 들면, 입력층(110)은 복수의 입력 노드들을 포함할 수 있다. 각 은닉층(121~12n)은 복수의 은닉 노드들을 포함할 수 있다. 출력층(130)은 복수의 출력 노드들을 포함할 수 있다. 각 층(입력층, 은닉층, 출력층)에 포함된 노드의 수는 서로 다를 수 있다. 입력 노드들은 각각 입력값들(IN1~INk)을 수신할 수 있다. 즉, 하나의 노드에는 하나의 입력값이 입력될 수 있다. 출력 노드들은 각각 출력값들(OUT1~OUT3)을 출력할 수 있다. 즉, 하나의 출력 노드에서 하나의 출력값이 출력될 수 있다. 하나의 노드와 다음 노드 사이에서 연결 가중치가 적용될 수 있다. 연결 가중치는 노드와 노드 사이에서 개별적으로 결정될 수 있다. 연결 가중치는 학습과 오류역전파 과정을 통해 업데이트될 수

있다.

[0036] 인공신경망 시스템(100)은 복수의 은닉층들(120)을 포함할 수 있다. 복수의 은닉층들(120)을 포함하는 인공신경망을 DNN(deep neural network)이라고 하고, DNN을 학습시키는 것을 딥러닝(deep learning)이라고 한다. 은닉층에 포함된 노드를 은닉 노드(hidden node)라고 한다. 이전 시간 구간에서의 은닉 노드의 출력은 현재 시간 구간에서의 은닉 노드들에 연결될 수 있다. 그리고, 현재 시간 구간에서의 은닉 노드의 출력은 다음 시간 구간에서의 은닉 노드들에 연결될 수 있다. 서로 다른 시간 구간에서 은닉 노드들 간에 재귀적(recurrent)인 연결이 있는 인공신경망을 RNN(recurrent neural network)이라고 한다.

[0037] 인공신경망 시스템(100)은 감독 학습(supervised learning)을 통해 신경망을 학습시킬 수 있다. 감독 학습이란 학습 데이터와 그에 대응하는 출력 데이터를 함께 신경망에 입력하고, 학습 데이터에 대응하는 출력 데이터가 출력되도록 연결선들의 연결 가중치를 업데이트하는 방법이다. 예를 들면, 인공신경망 시스템(100)은 델타 규칙(delta rule)과 오류역전파 학습(backpropagation learning) 등을 통해 인공 뉴런들 사이의 연결 가중치를 업데이트할 수 있다.

[0038] 오류역전파 학습은, 주어진 학습 데이터에 대해 전방 계산(forward computation)으로 오류를 추정한 후, 출력층에서 시작하여 은닉층과 입력층 방향으로 역으로 전진하여 추정한 오류를 전파하고, 오류를 줄이는 방향으로 연결 가중치를 업데이트하는 방법이다. 예를 들면, 오류역전파 학습은 입력(IN)을 각 층을 전달하여 출력(OUT)을 계산하는 단계 및 출력(OUT)의 오차를 줄이기 위해 신경망 내의 연결 가중치를 수정하는 단계로 구분될 수 있다.

[0039] 인공신경망 시스템(100)은 현재 설정된 연결 가중치들이 얼마나 최적에 가까운지를 측정하기 위한 목적 함수(objective function)를 정의하고, 목적 함수의 결과에 기초하여 연결 가중치들을 계속 변경하고, 학습을 반복적으로 수행할 수 있다. 예를 들면, 목적 함수는 신경망이 학습 데이터에 기초하여 실제 출력한 출력 값과 출력되기로 원하는 기대값 간의 오류를 계산하기 위한 오류 함수일 수 있다. 인공신경망 시스템(100)은 오류 함수의 값을 줄이는 방향으로 연결 가중치들을 업데이트할 수 있다.

[0040] 예를 들면, 연결 가중치를 업데이트하는 방법으로 경사 하강법(gradient descent)이 있다. 경사 하강법은 각 노드의 연결 가중치의 변화량에 따른 오류의 변화량(기울기)을 줄이는 방향으로 연결 가중치를 업데이트하는 것을 말한다. 인공신경망은 복수의 층으로 이루어져 있기 때문에 각 노드의 연결 가중치는 신호에 곱해지는 여러 연결 가중치 중에 하나이다. 따라서, 이러한 관계는 미적분의 연쇄 법칙(chain rule)을 통해 수학식 1과 같이 표현될 수 있다. 여기서, Error는 오류, weight는 연결 가중치, activation은 활성화 함수를 의미한다.

수학식 1

[0041]
$$\frac{dError}{dweight} = \frac{dError}{dactivation} * \frac{dactivation}{dweight}$$

[0042] 도 3은 도 2의 하나의 은닉 노드를 보여주는 도면이다. 도 3을 참조하면, 노드(ND)는 노드 입력들(NDIN)을 수신할 수 있다. 노드 입력들(NDIN)은 입력값들(X0~Xm)을 포함할 수 있다. 입력값들(X0~Xm)은 이전 은닉층의 노드들로부터 수신될 수 있다. 입력값들(X0~Xm)은 각각 연결 가중치(W0~Wm)와 곱해질 수 있다. 가중치(W0~Wm)는 이전 은닉층의 노드들과의 연결선과 일대일로 대응할 수 있다. 입력값들(X0~Xm) 각각과 연결 가중치(W0~Wm) 각각의 곱은 전달 함수(TF)에 의해 합해질 수 있다. 활성화 함수(AF)는 전달 함수(TF)의 출력값의 범위의 제한하여 노드 출력(NOUT)을 출력할 수 있다.

[0043] 활성화 함수(AF)는 전달 함수(TF)의 출력값의 범위를 제한하고 계산의 편의성을 제공할 수 있다. 따라서, 어떠한 활성화 함수(AF)를 사용하는지에 따라 인공신경망 시스템(100)의 학습 효율성이 달라질 수 있다.

[0044] 도 4는 일반적인 활성화 함수인 ReLU(Rectified Linear Unit)를 보여주는 도면이다. 도 4를 참조하면, ReLU는 입력값(즉, 전달 함수의 출력)이 0보다 작으면 0을 출력하고, 입력값이 0보다 크거나 같으면 입력값을 그대로 출력할 수 있다.

[0045] ReLU는 기존의 시그모이드(sigmoid) 함수의 문제점을 해결하기 위해 고안되었다. 활성화 함수로 시그모이드 함수를 사용하는 경우, 은닉층의 개수가 증가할수록 학습이 정상적으로 이루어지지 않는 현상(Vanishing

Gradient)이 발생하였다. 이러한 현상이 일어나는 이유는 시그모이드 함수는 출력값이 0~1 사이의 값을 갖도록 하는 함수이기 때문에 은닉층이 많아질수록 미분 값을 곱한 결과가 0에 근접하게 되어 입력층까지 오류가 역전파되지 않는다. 따라서, 이러한 문제점을 해결하기 위해 ReLU는 고안되었다. 다만, 최근 인공신경망 시스템(100)에 사용되는 고집적, 저전력 소자들은 채널 간섭 현상 및 불안정한 다층 구조로 인하여 소자 오류가 발생할 수 있고, 이에 따라 은닉층에서 매우 큰 전달 함수의 출력값이 생성될 수 있다. 따라서, 이러한 소자 오류에 의한 오류 전파 현상을 감소시키기 위한 활성화 함수가 필요하다.

[0046] 도 5는 본 발명의 실시 예에 따른 활성화 함수를 보여주는 도면이다. 도 5를 참조하면, 활성화 함수(AF2)는 수학식 2와 같이 입력값(x)이 0보다 큰 경우 2개의 구간에서 서로 다른 기울기(p, q)를 가지는 구간 선형 함수로 정의될 수 있다. 이하에서, 이러한 구간 선형 함수를 구간 선형 정류 유닛(Piecewise Linear Rectification Unit, PLRU)으로 지칭한다.

수학식 2

[0047]
$$g(x)=\begin{cases}0, & x<0\\ px, & a>x\geq 0\\ qx+a(p-q), & x\geq a\end{cases}$$

[0048] 입력값(x)이 음수인 경우, 활성화 함수(AF2)는 0을 출력할 수 있다. 입력값(x)이 0 이상이고 제1 파라미터(a)보다 작은 경우, 활성화 함수(AF2)는 제1 기울기(p)를 기울기로 가지고 원점을 지나는 직선에 따라 출력값(y)을 출력할 수 있다. 입력값(x)이 제1 파라미터(a)보다 큰 경우, 활성화 함수(AF2)는 제2 기울기(q)를 기울기로 가지고 [a, ap]인 점을 지나는 직선에 따라 출력값(y)을 출력할 수 있다. 제2 기울기(q)는 제1 기울기(p)보다 작은 값으로 결정될 수 있다. 따라서, 소자 오류로 인해 매우 큰 입력값이 전달되어도, 활성화 함수(AF2)는 출력값(y)의 범위를 제한할 수 있다. 제1 파라미터(a)는 0부터 단계적으로 증가시키면서 인공신경망 시스템(100)의 학습을 반복 수행하여 오류가 최소로 되는 지점으로 결정될 수 있다. 제1 기울기(p) 및 제2 기울기(q)는 제1 파라미터(a)가 결정된 후 오류역전파 학습 시 경사 하강법을 이용하여 결정할 수 있다.

[0049] 본 발명의 인공신경망 시스템(100)은 도 5에 설명된 바와 같이 구간별로 기울기가 감소되는 구간 선형 정류 유닛(PLRU)을 활성화 함수로서 사용할 수 있다. 따라서, 소자 오류로 인해 매우 큰 값을 전달하는 은닉 노드가 발생하여도, 인공신경망 시스템(100)은 활성화 함수를 통해 오류의 전파를 감소시킬 수 있다.

[0050] 도 6은 본 발명의 다른 실시 예에 따른 활성화 함수를 보여주는 도면이다. 도 6을 참조하면, 활성화 함수(AF3)는 수학식 3과 같이 입력값(x)이 0보다 큰 경우 3개의 구간에서 서로 다른 기울기(p, q, r)를 가지는 구간 선형 정류 유닛(PLRU)으로 정의될 수 있다.

수학식 3

[0051]
$$h(x)=\begin{cases}0, & x<0\\ px, & a>x\geq 0\\ qx+a(p-q), & b>x\geq a\\ rx+a(p-q)+b(q-r), & x\geq b\end{cases}$$

[0052] 입력값(x)이 음수인 경우, 활성화 함수(AF3)는 0을 출력할 수 있다. 입력값(x)이 0 이상이고 제1 파라미터(a)보다 작은 경우, 활성화 함수(AF3)는 제1 기울기(p)를 기울기로 가지고 원점을 지나는 직선에 따라 출력값(y)을 출력할 수 있다. 입력값(x)이 제1 파라미터(a) 이상이고 제2 파라미터(b)보다 작은 경우, 활성화 함수(AF3)는 제2 기울기(q)를 기울기로 가지고 [a, ap]인 점을 지나는 직선에 따라 출력값(y)을 출력할 수 있다. 입력값(x)이 제2 파라미터(b)보다 큰 경우, 활성화 함수(AF3)는 제3 기울기(r)를 기울기로 가지고 [b, a(p-q)+qb]인 점을 지나는 직선에 따라 출력값(y)을 출력할 수 있다.

[0053] 제2 기울기(q)는 제1 기울기(p)보다 작은 값으로 결정될 수 있다. 또한, 제3 기울기(r)는 제2 기울기(q)보다 작은 값으로 결정될 수 있다. 따라서, 소자 오류로 인해 매우 큰 입력값이 전달되어도, 활성화 함수(AF3)는 출력값(y)의 범위를 제한할 수 있다. 제1 파라미터(a)는 0부터 단계적으로 증가시키면서 인공신경망 시스템(100)의 학습을 반복 수행하여 오류가 최소로 되는 지점으로 결정될 수 있다. 제2 파라미터(b)는 제1 파라미터(a)부터 단계적으로 증가시키면서 인공신경망 시스템(100)의 학습을 반복 수행하여 오류가 최소로 되는 지점으로 결정될 수 있다. 제1 기울기(p), 제2 기울기(q) 및 제3 기울기(r)는 제1 파라미터(a) 및 제2 파라미터(b)가 결정된 후 오류역전파 학습 시 경사 하강법을 이용하여 결정할 수 있다.

[0054] 본 발명의 인공신경망 시스템(100)은 도 6에 설명된 바와 같이 구간별로 기울기가 감소되는 구간 선형 정류 유닛(PLRU)을 활성화 함수로서 사용할 수 있다. 따라서, 소자 오류로 인해 매우 큰 값을 전달하는 은닉 노드가 발생하여도, 인공신경망 시스템(100)은 활성화 함수를 통해 오류의 전파를 감소시킬 수 있다. 도 6의 구간 선형 정류 유닛(PLRU)은 도 5의 구간 선형 정류 유닛(PLRU)보다 오류의 전파를 더욱 감소시킬 수 있다.

[0055] 한편, 구간 선형 정류 유닛(PLRU)은 도 5 및 도 6에 한정되지 않고 입력값(x)이 0보다 큰 경우 복수의 구간들(4개 이상)로 구분되고, 각 구간은 선형 함수로 구성될 수 있다. 각 구간의 선형 함수는 구간 경계에서 서로 동일한 점을 지날 수 있다. 각 구간의 선형 함수의 기울기는 원점에서 멀어질수록 감소할 수 있다.

[0056] 이상에서와 같이 도면과 명세서에서 실시 예가 개시되었다. 여기서 특정한 용어들이 사용되었으나, 이는 단지 본 발명을 설명하기 위한 목적에서 사용된 것이지 의미 한정이나 특허 청구범위에 기재된 본 발명의 범위를 제한하기 위하여 사용된 것은 아니다. 그러므로 본 기술분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시 예가 가능하다는 점을 이해할 것이다. 따라서 본 발명의 진정한 기술적 보호 범위는 첨부된 특허 청구범위의 기술적 사상에 의해 정해져야 할 것이다.

부호의 설명

[0057] 100: 인공신경망 시스템

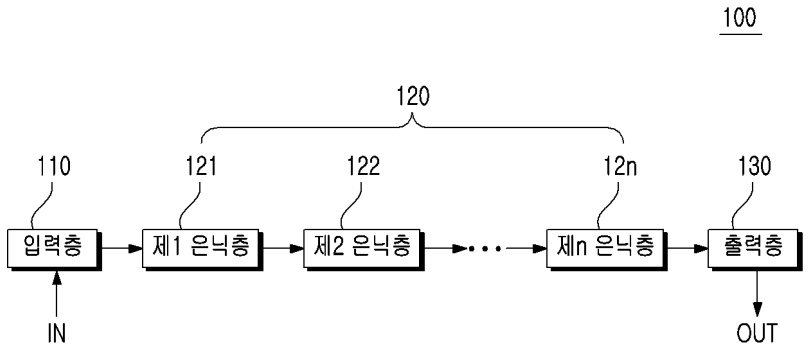
110: 입력층

120: 은닉층들

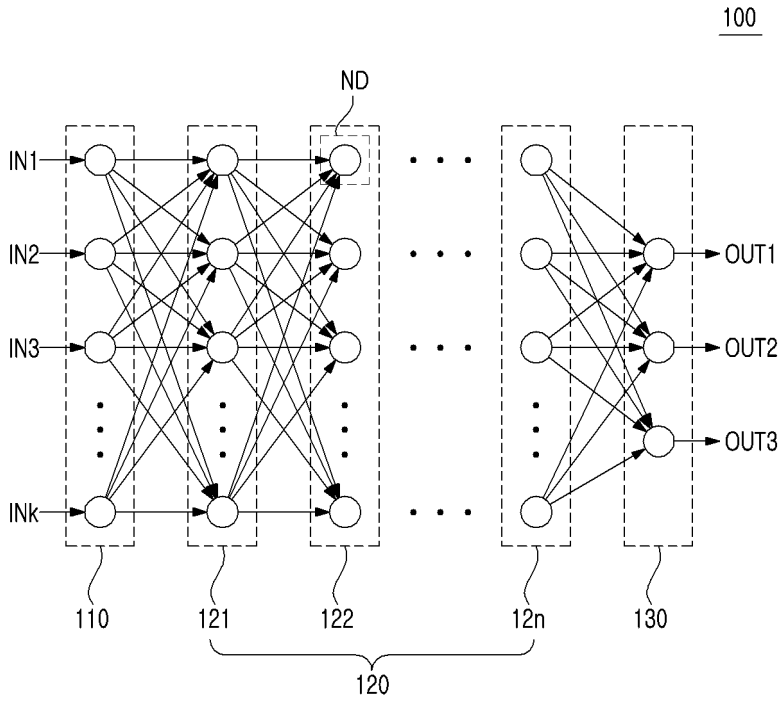
130: 출력층

도면

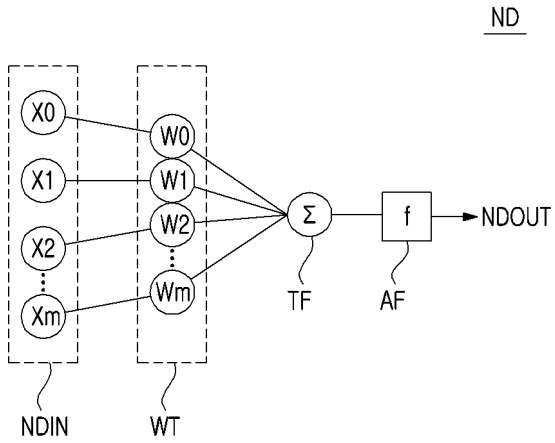
도면1



도면2

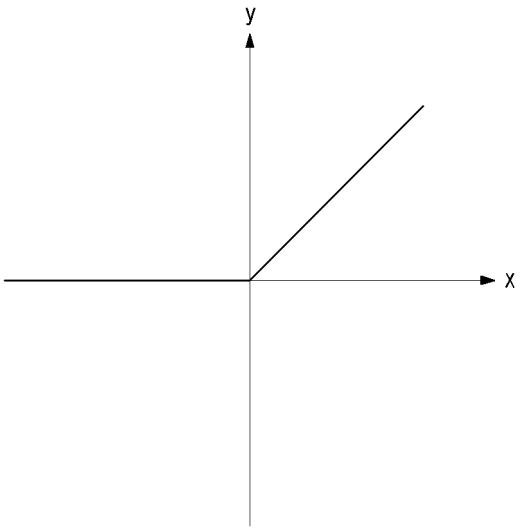

100
ND
IN1
IN2
IN3
INk
OUT1
OUT2
OUT3
110
121
122
12n
130
120

도면3


ND
X0
X1
X2
Xm
W0
W1
W2
Wm
Σ
f
NDOUT
TF
AF
NDIN
WT

도면4

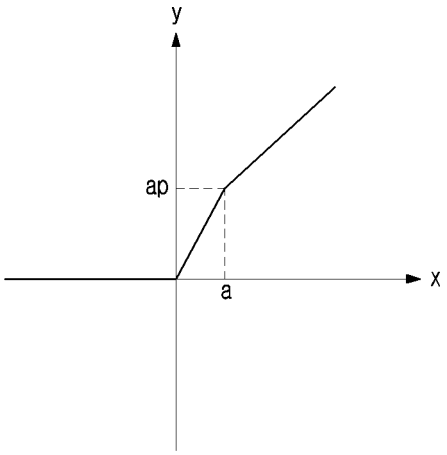
AF1



$$f(x) = \begin{cases} 0, & x < 0 \\ x, & x \geq 0 \end{cases}$$

도면5

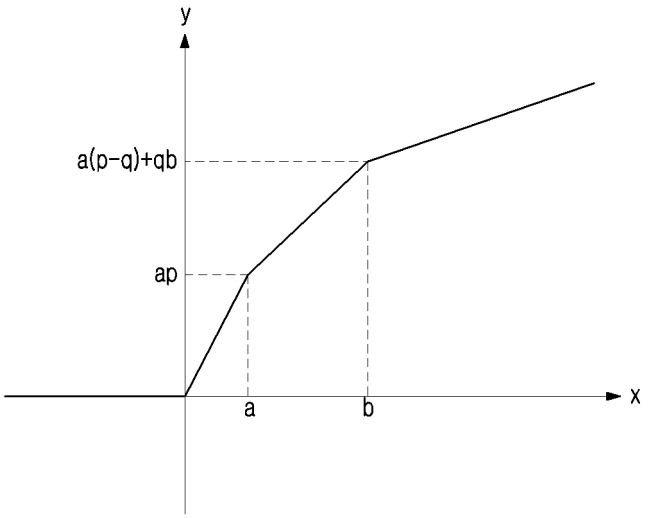
AF2



$$g(x) = \begin{cases} 0, & x < 0 \\ px, & a > x \geq 0 \\ qx+a(p-q), & x \geq a \end{cases}$$

도면6

AF3



$$h(x) = \begin{cases} 0, & x < 0 \\ px, & a > x \geq 0 \\ qx + a(p-q), & b > x \geq a \\ rx + a(p-q) + b(q-r), & x \geq b \end{cases}$$