



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0082166
(43) 공개일자 2024년06월10일

(51) 국제특허분류(Int. Cl.)
G06N 3/09 (2023.01) G06F 17/16 (2006.01)
G06N 3/048 (2023.01)
(52) CPC특허분류
G06N 3/09 (2023.01)
G06F 17/16 (2013.01)
(21) 출원번호 10-2023-0100210
(22) 출원일자 2023년08월01일
심사청구일자 2023년08월01일
(30) 우선권주장
1020220163786 2022년11월30일 대한민국(KR)

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암
동5가)
(72) 발명자
김미정
서울특별시 종로구 비봉1길 22-4 삼진파크맨션 나
동 102호
(74) 대리인
특허법인위솔

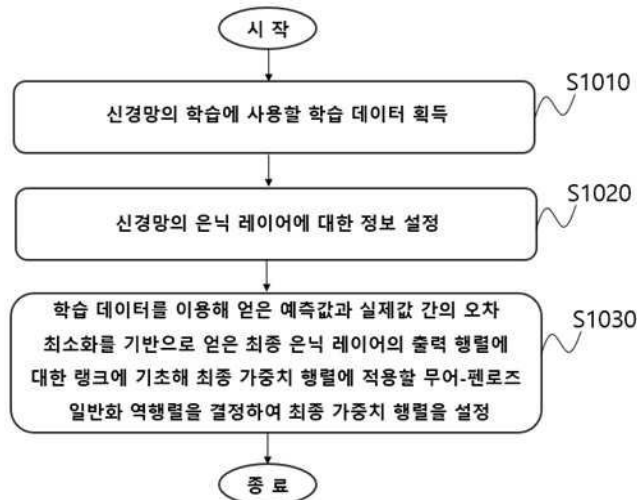
전체 청구항 수 : 총 10 항

(54) 발명의 명칭 무어-펜로즈 일반화 역행렬 기반의 신경망 신경망 생성 방법 및 이를 위한 장치

(57) 요약

일 실시 예에 따른 신경망 생성 장치는 하나 이상의 프로세서에 의해, 신경망의 학습에 사용할 학습 데이터를 획득하는 동작; 신경망의 은닉 레이어에 대한 정보를 설정하는 동작; 및 학습 데이터를 이용해 얻은 예측값과 실제값 간의 오차 값을 최소화할 기반으로 얻은 은닉 레이어 중 최종 레이어의 출력 행렬에 대한 랭크에 기초해 최종 가중치 행렬에 적용할 무어-펜로즈 일반화 역행렬을 결정하여 최종 가중치 행렬을 설정하는 동작을 수행할 수 있다.

대표도 - 도2



(52) CPC특허분류

G06N 3/048 (2023.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711164829
과제번호	2019R1A2C1002706
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	개인기초연구(과기정통부)
연구과제명	5세대 네트워크 기술을 위한 데이터 트래픽 모델링및 예측 기술 연구: 머신러닝 및
시계열 분석 접근법	
기 여 율	1/1
과제수행기관명	고려대학교
연구기간	2022.03.01 ~ 2023.02.28

명세서

청구범위

청구항 1

프로세서 및 메모리를 포함하는 장치가 무어-펜로즈 일반화 역행렬 기반의 친환경 신경망을 생성하는 방법에 있어서,

신경망의 학습에 사용할 학습 데이터를 획득하는 동작;

상기 신경망의 은닉 레이어에 대한 정보를 설정하는 동작; 및

상기 학습 데이터를 이용해 얻은 예측값과 실제값 간의 오차 최소화를 기반으로 얻은 상기 은닉 레이어 중 최종 레이어의 출력 행렬에 대한 랭크에 기초해 최종 가중치 행렬에 적용할 무어-펜로즈 일반화 역행렬을 결정하여 상기 최종 가중치 행렬을 설정하는 동작;

을 포함하는 방법.

청구항 2

제1항에 있어서,

상기 학습 데이터 중 로우 데이터가 $\{X_i\}_{i=1}^m$ (m 은 로우 데이터의 개수)이고, 각 로우 데이터의 특징을 n 개로

설정하여, i 번째 로우 데이터 X_i 에 대한 각 특징값의 표현을 $X_i = (X_{i,1}, L, X_{i,n})$ 로 할 때,

상기 학습 데이터를 획득하는 동작은

X_i 에 대한 각 특징값을 기초로 $X_i = (X_{i,1}, L, X_{i,n})$ 에 대해 라벨 Y_i 을 할당하고, 로우 데이터

X_i 에 라벨 Y_i 를 매핑하여, 라벨이 매핑된 $\{(X, Y)\} = \{(X_i, Y_i)\}_{i=1}^m$ 데이터 집합을 생성하는 동작을 포함하는,

방법.

청구항 3

제2항에 있어서,

상기 학습 데이터를 획득하는 동작은

상기 데이터 집합 $\{(X, Y)\} = \{(X_i, Y_i)\}_{i=1}^m$ 에 대해 하기 수학식 1을 기초로 min-max 변환을 수행하여, 학

습 집합 $\{(x_i, y_i)\}_{i=1}^{m_1}$ ($m_1 = m \times r$, r 은 상기 데이터 집합에서 학습 집합에 사용될 비율,

$r \in (0, 1)$)을 생성하는 동작을 포함하는,

[수학식 1]

$$\text{min-max}(x_{ij}) = \frac{x_{ij} - \min(\mathbf{x}^i)}{\max(\mathbf{x}^i) - \min(\mathbf{x}^i)}, j = 1, L, n,$$

($\mathbf{x}^i$ 는 $\{\mathbf{x}_i\}_{i=1}^m$ 의 i 번째 특징으로 이루어진 열벡터)

방법.

청구항 4

제1항에 있어서,

상기 은닉 레이어에 대한 정보를 설정하는 동작은

은닉 레이어의 활성화 함수 f , 각 은닉 레이어의 노드 개수 k_0 및 $\{k_p\}_{p=1}^P$ (p 는 은닉 레이어를 특징하는 인덱스)를 포함하는

방법.

청구항 5

제1항에 있어서,

상기 은닉 레이어에 대한 정보를 설정하는 동작은

상기 최종 은닉 레이어와 출력 레이어 사이의 가중치 행렬의 랭크 $r(W^P)$ 가 최대가 되도록, 은닉 레이어의 가중치 행렬 및 바이어스 행렬 $\{W^p, b^p\}_{p=1}^P$ 의 초기값을 설정하는 동작을 포함하는, 방법.

청구항 6

제5항에 있어서,

상기 은닉 레이어에 대한 정보를 설정하는 동작은

입력 레이어로부터 출력 레이어로 향하는 방향으로 각 은닉 레이어에 대해, p 번째(p 는 은닉 레이어를 특징하는

인덱스) 레이어의 출력 함수 H_p 의 랭크 $r(H_p)$ 가 p 번째 레이어의 출력 함수의 행의 개수 또는 열의 개수 중의 최소값 $\min(k_{p-1}, k_p)$ 과 같든지 판별하는 동작; 및

상기 판별에서 $r(H_p) = \min(k_{p-1}, k_p)$ 를 만족하지 않는다면, 기 설정된 횟수만큼 p 번째 레이어의 바이어

스를 갱신하여 p 번째 레이어의 출력 함수 H_p 를 갱신하는 동작을 포함하는,

방법.

청구항 7

제6항에 있어서,

상기 최종 가중치 행렬을 설정하는 동작은

상기 학습 데이터의 예측값과 상기 학습 데이터의 실제값의 L2 norm 오차 E_2 가 최소가 되는 상기 최종 레이어의

출력 행렬 H_P 을 결정하고, 상기 출력 행렬에 대한 랭크 $r(H_P)$ 에 따라 상기 최종 가중치 행렬에 적용

할 무어-펜로즈 일반화 역행렬 H_P^{r-} , $(H_P^T H_P)^{-1}$, $(H_P^T H_P)^+$ 를 결정하여, 상기 결정에 따라 상기 최종

가중치 행렬 W^{out} (E_2 가 0일 때) 또는 W^{apout} (E_2 가 0이 아닐 때)를 설정하는 동작을 포함하는,

[수학식 2]

$$\left\{ \begin{array}{l} \text{(i) } r(H_P) = k_{p-1} \text{인 경우: } W^{out} = H_P^r y^o \text{이고 } E_2 \equiv 0 \\ \text{(ii) } r(H_P) = k_p \text{인 경우} \\ \quad \text{(a) } y \in C(H_P): W^{out} = (H_P^T H_P)^{-1} H_P^T y^o \text{이고 } E_2 \equiv 0 \\ \quad \text{(b) } y \notin C(H_P): W^{apout} = (H_P^T H_P)^{-1} H_P^T y^o \text{일 때 } E_2 \text{는 최소값} \\ \text{(iii) } r(H_P) < \min(k_{p-1}, k_p) \text{인 경우} \\ \quad \text{(a) } y \in C(H_P): W^{out} \text{의 성분 중에는 자유변수가 있고 } E_2 \equiv 0 \\ \quad \text{(b) } y \notin C(H_P): W^{apout} = (H_P^T H_P)^+ H_P^T y^o \text{일 때 } E_2 \text{는 최소값} \end{array} \right.$$

(W^{apout} 는 W^{out} 의 근사행렬, $C(H_P)$ 는 열공간(column space), (iii)-(a)에서 W^{out} 의 성분 중

자유변수(free variables)의 개수는 $\min(k_{p-1}, k_p) - r(H_P)$)

방법.

청구항 8

제1항에 있어서,

상기 신경망은

ELM(Extreme Learning Machine) 또는 RVFL(Random Vector Functional Link)을 기초로 네트워크가 설계되는

방법.

청구항 9

명령어를 포함하는 메모리; 및

상기 명령어를 실행하는 프로세서를 포함하고,

상기 프로세서의 동작은,

신경망의 학습에 사용할 학습 데이터를 획득하는 동작;

상기 신경망의 은닉 레이어에 대한 정보를 설정하는 동작; 및

상기 신경망의 최종 은닉 레이어의 출력 행렬에 대한 랭크에 기초해 상기 최종 은닉 레이어와 출력 레이어 사이의 가중치 행렬에 적용할 무어-펜로즈 일반화 역행렬을 결정하는 동작;

을 포함하는 장치.

청구항 10

컴퓨터 판독 가능 기록매체에 저장되어 있는 컴퓨터 프로그램으로서,

상기 컴퓨터 프로그램은, 프로세서에 의해 실행되면,

신경망의 학습에 사용할 학습 데이터를 획득하는 동작;

상기 신경망의 은닉 레이어에 대한 정보를 설정하는 동작; 및

상기 학습 데이터를 이용해 얻은 예측값과 실제값 간의 오차 최소화를 기반으로 얻은 상기 은닉 레이어 중 최종 레이어의 출력 행렬에 대한 랭크에 기초해 최종 가중치 행렬에 적용할 무어-펜로즈 일반화 역행렬을 결정하여 상기 최종 가중치 행렬을 설정하는 동작;

을 포함하는 방법을 상기 프로세서가 수행하도록 하기 위한 명령어를 포함하는 컴퓨터로 판독 가능한 매체에 저장된 컴퓨터 프로그램.

발명의 설명

기술 분야

[0001] 본 발명은 무어-펜로즈 일반화 역행렬 기반의 친환경 신경망 생성 방법 및 이를 위한 장치에 관한 것이다. 보다 자세하게는 신경망의 가중치를 무어-펜로즈 일반화 역행렬을 기반으로 결정하여 친환경적으로 신경망을 생성하는 방법 및 이를 위한 장치에 관한 것이다.

배경 기술

[0003] 최근 딥러닝은 컴퓨터 네트워크, 의학, 의료, 기상학, 경제학 등 다양한 분야에서 발생하는 예측 문제를 다루는데 활용되고 있다. 이러한 딥러닝을 이용한 예측은 일반적으로 좋은 성능을 보이지만 그 메커니즘을 분석하여 성능을 설명하는 것은 어려운 것으로 간주되어 왔다.

[0004] 한편, 딥러닝에 광범위하게 사용되고 있는 학습 알고리즘은 주로 가중치 및 바이어스를 반복적으로 갱신하여 학습 속도가 느린 경사 하강법 알고리즘을 기반으로 하고 있다.

[0005] 다만, 경사 하강법은 은닉층에서 최적의 가중치 및 바이어스를 찾거나 원하는 성능을 얻기까지 막대한 횟수의 반복 학습이 필수적이기 때문에, 많은 시간과 자원이 소모되며, 경우에 따라 경사 하강법에 따른 가중치 갱신 도중 국지적 최소값에 수렴하여 성능 개선에 문제가 생기는 등의 단점이 있다.

[0006] 이러한 단점을 극복하기 위해 지난 수십년 동안 다양한 이론적 접근이 고려되어 왔다.

선행기술문헌

특허문헌

[0008] (특허문헌 0001) 대한민국 등록특허공보 제10-2152615호(2020.09.08)

발명의 내용

해결하려는 과제

[0009] 본 발명이 해결하고자 하는 기술적 과제는 다양한 설계에 따른 신경망의 하이퍼파라미터 설정에 있어서, 무어-펜로즈 일반화 역행렬(Moore-Penrose generalized inverse)의 극한을 이용하여 신경망의 설계에 맞는 최적의 하이퍼파라미터를 곧바로 적용함으로써, 가중치 학습에 소모되는 자원을 획기적으로 감소시키는 친환경적인 신경망을 생성하는 기술을 제공하는 것이다.

[0010] 본 발명의 기술적 과제들은 이상에서 언급한 기술적 과제들로 제한되지 않으며, 언급되지 않은 또 다른 기술적 과제들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

과제의 해결 수단

[0012] 상기 기술적 과제를 달성하기 위한 본 발명의 일 실시 예에 따른 프로세서 및 메모리를 포함하는 장치가 무어-펜로즈 일반화 역행렬 기반의 친환경 신경망을 생성하는 방법은 신경망의 학습에 사용할 학습 데이터를 획득하는 동작; 상기 신경망의 은닉 레이어에 대한 정보를 설정하는 동작; 및 상기 학습 데이터를 이용해 얻은 예측값과 실제값 간의 오차 최소화를 기반으로 얻은 상기 은닉 레이어 중 최종 레이어의 출력 행렬에 대한 랭크에 기초해 최종 가중치 행렬에 적용할 무어-펜로즈 일반화 역행렬을 결정하여 상기 최종 가중치 행렬을 설정하는 동작을 포함할 수 있다.

[0013] 일 실시 예에 따르면, 상기 학습 데이터 중 로우 데이터(raw data)가 $\{X_i\}_{i=1}^m$ (m 은 로우 데이터의 개수)이고, 각 로우 데이터의 특징을 n 개로 설정하여, i 번째 로우 데이터 X_i 에 대한 각 특징값의 표현을 $X_i = (X_{i,1}, L, X_{i,n})$ 로 할 때, 상기 학습 데이터를 획득하는 동작은 X_i 에 대한 각 특징값을 기초로 $X_i = (X_{i,1}, L, X_{i,n})$ 에 대해 라벨 Y_i 을 할당하고, 로우 데이터 X_i 에 라벨 Y_i 를 매핑하여, 라벨이 매핑된 $\{(X, Y)\} = \{(X_i, Y_i)\}_{i=1}^m$ 데이터 집합을 생성하는 동작을 포함할 수 있다.

[0014] 일 실시 예에 따르면, 상기 학습 데이터를 획득하는 동작은 상기 데이터 집합 $\{(X, Y)\} = \{(X_i, Y_i)\}_{i=1}^m$ 에 대해 하기 수학식 1을 기초로 min-max 변환을 수행하여, 학습 집합 $\{(x_i, y_i)\}_{i=1}^{m_1}$ ($m_1 = m \times r$, r 은 상기 데이터 집합에서 학습 집합에 사용될 비율, $r \in (0, 1)$)을 생성하는 동작을 포함할 수 있다.

[0015] [수학식 1]

$$\min\text{-max}(x_j) = \frac{x_j - \min(\mathbf{x}^i)}{\max(\mathbf{x}^i) - \min(\mathbf{x}^i)}, j = 1, L, n,$$

[0016]

[0017] ($\mathbf{x}^i$ 는 $\{\mathbf{x}_i\}_{i=1}^m$ $\{\mathbf{x}_i\}_{i=1}^m$ 의 i 번째 특징으로 이루어진 열벡터)

[0018] 일 실시 예에 따르면, 상기 은닉 레이어에 대한 정보를 설정하는 동작은 은닉 레이어의 활성화 함수 f , 각 은닉 레이어의 노드 개수 k_0 및 $\{k_p\}_{p=1}^P$ (p 는 은닉 레이어를 특징하는 인덱스)를 포함할 수 있다.

[0019] 일 실시 예에 따르면, 상기 은닉 레이어에 대한 정보를 설정하는 동작은 상기 최종 은닉 레이어와 출력 레이어 사이의 가중치 행렬의 랭크 $r(W^P)$ 가 최대가 되도록, 은닉 레이어의 가중치 행렬 및 바이어스 행렬 $\{W^p, b^p\}_{p=1}^P$ 의 초기값을 설정하는 동작을 포함할 수 있다.

[0020] 일 실시 예에 따르면, 상기 은닉 레이어에 대한 정보를 설정하는 동작은 입력 레이어로부터 출력 레이어로 향하는 방향으로 각 은닉 레이어에 대해, p 번째(p 는 은닉 레이어를 특징하는 인덱스) 레이어의 출력 함수 H_p 의 랭크 $r(H_p)$ 가 p 번째 레이어의 출력 함수의 행의 개수 또는 열의 개수 중의 최소값 $\min(k_{p-1}, k_p)$ 과 같은지 판별하는 동작; 및 상기 판별에서 $r(H_p) = \min(k_{p-1}, k_p)$ 를 만족하지 않는다면, 기 설정된 횟수만큼 p 번째 레이어의 바이어스를 갱신하여 p 번째 레이어의 출력 함수 H_p 를 갱신하는 동작을 포함할 수 있다.

[0021] 일 실시 예에 따르면, 상기 최종 가중치 행렬을 설정하는 동작은 상기 최종 레이어의 출력 행렬 H_P 에 대한 랭크 $r(H_P)$ 와, 학습과정 중에 상기 학습 데이터에 대한 예측값과 상기 학습 데이터의 실제값의 L2 norm 오차가 최소가 되도록 하는 최종 가중치 행렬을 결정하고, 최종 가중치 행렬의 랭크에 따라 적용할 무어-펜로즈 일반화 역행렬 $H_P^{r-}, (H_P^T H_P)^{-1}, (H_P^T H_P)^{+}$ 을 결정한다. 그 결정 결과에 따라 하기 수학적 식 2에 나타난 바와 같이 데이터의 예측값과 실제값의 L2 norm 오차 E_2 가 0 인지 또는 0보다 큰 실수인지가 구분된다. 상기 결정에 따라 상기 최종 가중치 행렬 W^{out} (E_2 를 0으로 줌) 또는 W^{apout} (E_2 를 0보다 큰 실수로 줌)을 설정하는 동작을 포함할 수 있다. 결정된 최종 가중치 행렬을 테스트 데이터에 적용하여 가중치 행렬의 타당성을 검증한다.

[0022] [학식 2]

[0023]
$$\left\{ \begin{array}{l} \text{(i) } r(H_P) = k_{P-1} \text{인 경우: } W^{out} = H_P^T y^o \text{이고 } E_2 \equiv 0 \\ \text{(ii) } r(H_P) = k_P \text{인 경우} \\ \quad \text{(a) } y \in C(H_P): W^{out} = (H_P^T H_P)^{-1} H_P^T y^o \text{이고 } E_2 \equiv 0 \\ \quad \text{(b) } y \notin C(H_P): W^{apout} = (H_P^T H_P)^{-1} H_P^T y^o \text{일 때 } E_2 \text{는 최소값} \\ \text{(iii) } r(H_P) < \min(k_{P-1}, k_P) \text{인 경우} \\ \quad \text{(a) } y \in C(H_P): W^{out} \text{의 성분 중에는 자유변수가 있고 } E_2 \equiv 0 \\ \quad \text{(b) } y \notin C(H_P): W^{apout} = (H_P^T H_P)^+ H_P^T y^o \text{일 때 } E_2 \text{는 최소값} \end{array} \right.$$

[0024] (W^{apout} 는 W^{out} 의 근사행렬, $C(H_p)$ 는 열공간(column space), (iii)-(a)에서 W^{out} 의 성분 $\min(k_{p-1}, k_p) - r(H_p)$) 중 자유변수(free variables)의 개수는

[0025] 일 실시 예에 따르면, 상기 신경망은 ELM(Extreme Learning Machine) 또는 RVFL(Random Vector Functional Link)을 기초로 네트워크가 설계될 수 있다.

[0026] 상기 기술적 과제를 달성하기 위한 본 발명의 일 실시 예에 따른 장치는 명령어를 포함하는 메모리; 및 상기 명령어를 실행하는 프로세서를 포함하고, 상기 프로세서의 동작은 신경망의 학습에 사용할 학습 데이터를 획득하는 동작; 상기 신경망의 은닉 레이어에 대한 정보를 설정하는 동작; 및 상기 신경망의 최종 은닉 레이어의 출력 행렬에 대한 랭크에 기초해 상기 최종 은닉 레이어와 출력 레이어 사이의 가중치 행렬에 적용할 무어-펜로즈 일반화 역행렬을 결정하는 동작을 포함할 수 있다.

[0027] 상기 기술적 과제를 달성하기 위한 본 발명의 일 실시 예에 따른 컴퓨터 판독 가능 기록매체에 저장되어 있는 컴퓨터 프로그램으로서, 상기 컴퓨터 프로그램은, 프로세서에 의해 실행되면, 신경망의 학습에 사용할 학습 데이터를 획득하는 동작; 상기 신경망의 은닉 레이어에 대한 정보를 설정하는 동작; 및 상기 학습 데이터를 사용한 학습을 통해 상기 은닉 레이어 중 최종 레이어의 출력 행렬에 대한 랭크에 기초해 최종 가중치 행렬에 적용할 무어-펜로즈 일반화 역행렬을 결정한다. 상기 최종 가중치 행렬을 설정하는 동작을 포함하는 방법을 상기 프로세서가 수행하도록 하기 위한 명령어를 포함할 수 있다. 결정된 최종 가중치 행렬을 테스트 데이터에 적용하여 가중치 행렬의 타당성을 검증하는 과정을 포함한다.

발명의 효과

[0029] 상기와 같은 본 발명에 따르면, 수학적 이론을 기반으로 학습 과정에서 파생된 행렬의 특성을 파악하고, 정확한 예측과 학습 속도에 중점을 둔 무어-펜로즈 일반화 역행렬(Moore-Penrose generalized inverse)을 이용한 극한 접근 방식을 적용함으로써, 가중치 학습에 소모되는 자원을 획기적으로 감소시키면서도 신경망 모델이 정확한 출력값을 도출하게 하여 예측 오차를 최소화할 수 있다.

[0030] 이를 통해, 본 발명은 시간과 자원이 많이 소요되는 기존의 가중치 학습 방법에 대한 단점을 극복하고, 학습에 적은 시간이 소요되면서도 높은 정확도의 예측이 가능할 뿐만 아니라 기존의 기술들에 비해 적은 자원을 사용하여 친환경적이다.

[0031] 또한, 본 발명의 기술은 다양한 신경망 모델의 가중치 설정에 사용될 수 있다. 예를 들면, 전자, 환경, 의료, 보건 등의 다양한 분야의 데이터와 신경망 모델에 적용하여 사용할 수 있다.

[0032] 본 발명의 효과들은 이상에서 언급한 효과들로 제한되지 않으며, 언급되지 않은 또 다른 효과들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

[0034] 도 1은 일 실시 예에 따른 신경망 생성 장치가 포함하는 전체 구성을 예시적으로 도시한 도면이다.

도 2는 일 실시 예에 따른 신경망 생성 방법의 대표적인 단계를 나타낸 순서도이다.

도 3은 일 실시 예에 따라 ELM(Extreme Learning Machine)을 기초로 설계한 신경망의 구조를 예시적으로 도시한 도면이다.

도 4는 일 실시 예에 따라 RVFL(Random Vector Functional Link)을 기초로 설계한 신경망의 구조를 예시적으로 도시한 도면이다.

도 5는 일 실시 예에 따른 신경망 생성 방법을 수학적 표현과 함께 예시적으로 도시한 도면이다.

발명을 실시하기 위한 구체적인 내용

- [0035] 본 발명의 목적과 기술적 구성 및 그에 따른 작용 효과에 관한 자세한 사항은 본 발명의 명세서에 첨부된 도면에 의거한 이하의 상세한 설명에 의해 보다 명확하게 이해될 것이다. 첨부된 도면을 참조하여 본 발명에 따른 실시 예를 상세하게 설명한다.
- [0036] 본 명세서에서 개시되는 실시 예들은 본 발명의 범위를 한정하는 것으로 해석되거나 이용되지 않아야 할 것이다. 이 분야의 통상의 기술자에게 본 명세서의 실시 예를 포함한 설명은 다양한 응용을 갖는다는 것이 당연하다. 따라서, 본 발명의 상세한 설명에 기재된 임의의 실시 예들은 본 발명을 보다 잘 설명하기 위한 예시적인 것이며 본 발명의 범위가 실시 예들로 한정되는 것을 의도하지 않는다.
- [0037] 도면에 표시되고 아래에 설명되는 기능 블록들은 가능한 구현의 예들일 뿐이다. 다른 구현들에서는 상세한 설명의 사상 및 범위를 벗어나지 않는 범위에서 다른 기능 블록들이 사용될 수 있다. 또한, 본 발명의 하나 이상의 기능 블록이 개별 블록들로 표시되지만, 본 발명의 기능 블록들 중 하나 이상은 동일 기능을 실행하는 다양한 하드웨어 및 소프트웨어 구성들의 조합일 수 있다.
- [0038] 또한, 어떤 구성요소들을 포함한다는 표현은 "개방형"의 표현으로서 해당 구성요소들이 존재하는 것을 단순히 지칭할 뿐이며, 추가적인 구성요소들을 배제하는 것으로 이해되어서는 안 된다.
- [0039] 나아가 어떤 구성요소가 다른 구성요소에 "연결되어" 있다거나 "접속되어" 있다고 언급될 때에는, 그 다른 구성요소에 직접적으로 연결 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 한다.
- [0040] 이하에서는 도면들을 참조하여 본 발명의 세부적인 실시 예들에 대해 살펴보도록 한다.
- [0041] 도 1은 일 실시 예에 따른 신경망 생성 장치(100)(이하, '장치(100)'로 지칭)가 포함하는 전체 구성을 예시적으로 도시한 도면이다.
- [0042] 그러나 이는 본 발명의 목적을 달성하기 위한 바람직한 실시 예일 뿐이며, 필요에 따라 일부 구성이 추가되거나 삭제될 수 있고, 어느 한 구성이 수행하는 역할을 다른 구성이 함께 수행할 수도 있음은 물론이다.
- [0043] 일 실시 예에 따른 장치(100)는 프로세서(10), 네트워크 인터페이스(20), 메모리(30), 스토리지(40) 및 이들을 연결하는 데이터 버스(50)를 포함할 수 있으며, 기타 본 발명의 목적을 달성함에 있어 요구되는 부가적인 구성들을 더 포함할 수 있음은 물론이라 할 것이다.
- [0044] 프로세서(10)는 각 구성의 전반적인 동작을 제어한다. 프로세서(10)는 CPU(Central Processing Unit), MPU(Micro Processor Unit), MCU(Micro Controller Unit) 또는 본 발명이 속하는 기술 분야에서 널리 알려져 있는 형태의 인공지능 프로세서 중 어느 하나일 수 있다. 아울러, 프로세서(10)는 일 실시 예에 따른 장치(100)가 수행하는 동작 또는 오퍼레이션을 수행하기 위한 적어도 하나의 애플리케이션 또는 프로그램에 대한 연산을 수행할 수 있다.
- [0045] 네트워크 인터페이스(20)는 일 실시 예에 따른 장치(100)의 유무선 인터넷 통신을 지원하며, 그 밖의 공지의 통신 방식을 지원할 수도 있다. 따라서 네트워크 인터페이스(20)는 그에 따른 통신 모듈을 포함하여 구성될 수 있다.
- [0046] 메모리(30)는 각종 정보, 명령 및/또는 정보를 저장하며, 일 실시 예에 따른 장치(100)가 수행하는 동작 또는 오퍼레이션을 수행하기 위해 스토리지(40)로부터 하나 이상의 컴퓨터 프로그램(41)을 로드할 수 있다. 도 1에서는 메모리(30)의 하나로 RAM을 도시하였으나 이와 더불어 다양한 저장 매체를 메모리(30)로 이용할 수 있음은 물론이다.
- [0047] 스토리지(40)는 하나 이상의 컴퓨터 프로그램(41)을 비임시적으로 저장할 수 있다. 이러한 스토리지(40)는 ROM(Read Only Memory), EPROM(Erasable Programmable ROM), EEPROM(Electrically Erasable Programmable ROM), 플래시 메모리 등과 같은 비휘발성 메모리, 하드 디스크(HDD), 보조 저장 매체(SSD), 착탈형 디스크, 또는 본 발명이 속하는 기술 분야에서 널리 알려져 있는 임의의 형태의 컴퓨터로 읽을 수 있는 기록 매체 중 어느 하나일 수 있다.
- [0048] 컴퓨터 프로그램(41)은 메모리(30)에 로드되어, 하나 이상의 프로세서(10)에 의해, 신경망의 학습에 사용할 학습 데이터를 획득하는 동작; 신경망의 은닉 레이어에 대한 정보를 설정하는 동작; 및 은닉 레이어 중 최종 레이어의 출력 행렬에 대한 랭크와 학습 데이터에 대한 오차값에 기초해 최종 가중치 행렬에 적용할 무어-웬로즈 일반화 역행렬을 결정하여 최종 가중치 행렬을 설정하는 동작 또는 오퍼레이션을 실행할 수 있다.

- [0049] 이상 간단하게 언급한 컴퓨터 프로그램(41)이 수행하는 동작 또는 오퍼레이션은 컴퓨터 프로그램(41)의 일 기능으로 볼 수 있으며, 보다 자세한 설명은 일 실시 예에 따른 장치(100)가 수행하는 방법에 대한 설명에서 후술하도록 한다.
- [0050] 데이터 버스(50)는 이상 설명한 프로세서(10), 네트워크 인터페이스(20), 메모리(30) 및 스토리지(40) 사이의 명령 및/또는 정보의 이동 경로가 된다.
- [0051] 이상 간단하게 설명한 일 실시 예에 따른 장치(100)는 독립된 디바이스의 형태, 예를 들어 전자 기기나 서버(클라우드 포함)의 형태일 수 있으며, 여기서 전자 기기는 한 장소에 고정 설치되어 사용하는 데스크톱 PC, 서버 디바이스 등과 같은 기기 뿐만 아니라, 스마트폰, 태블릿 PC, 노트북 PC, PDA, PMP 등과 같이 휴대가 용이한 포터블 기기 등이라도 무방한바, 프로세서(10)에 해당하는 CPU 등이 설치되고 네트워크 기능만 보유하고 있는 전자 기기라면 어떠한 것이라도 무방하다 할 것이다. 예를 들어, 일 실시 예에 따른 장치(100)는 독립된 디바이스인 전자 기기 형태 중 "서버"의 형태로 구현될 수 있다. 또한, 일 실시 예에 따른 장치(100)는 독립된 디바이스인 전자 기기 형태 중 서버가 제공하는 애플리케이션 또는 서비스를 이용하기 위해 사용자가 소지하는 "단말"의 형태로 구현될 수 있다.
- [0052] 이하, 일 실시 예에 따른 장치(100)가 수행하는 방법에 대하여 도 2 내지 도 5를 참조하여 설명하도록 한다.
- [0053] 도 2는 일 실시 예에 따른 신경망 생성 방법의 대표적인 단계를 나타낸 순서도이다.
- [0054] 도 2에 개시된 각 단계는 일 실시 예에 따른 장치(100)를 통해 수행될 수 있고, 구체적으로 각 단계의 동작은 프로세서(10)가 수행하는 동작으로 이해될 수 있다.
- [0055] 도 2에 개시된 각 단계는 본 발명의 목적을 달성함에 있어서 바람직한 실시 예일 뿐이며, 필요에 따라 일부 단계가 추가 또는 삭제될 수 있음은 물론이고, 어느 한 단계가 다른 단계에 포함되어 수행될 수도 있다.
- [0056] 도 2에 개시된 각 동작의 순서는 이해의 편의를 위해 배치된 순서일 뿐, 이러한 순서가 시계열적인 순서로 한정되는 것이 아니며, 설계자의 선택에 따라 순서가 다르게 변경되어 동작될 수 있다.
- [0057] S1010 단계에서, 장치(100)는 신경망의 학습에 사용할 학습 데이터를 획득할 수 있다.

[0058] 학습 데이터 중 로우 데이터를 $\{X_i\}_{i=1}^m$ (m 은 로우 데이터의 개수)로 표현하고, 각 로우 데이터의 특징을 n 개

로 특징하여 사용한다면, i 번째 로우 데이터 X_i 에 대한 각 특징값의 표현을 $X_i = (X_{i,1}, L, X_{i,n})$ 로 할 수 있다.

[0059] 또한, X_i 에 할당되는 라벨을 Y_i 로 표현하기로 할 때, 장치(100)는 X_i 에 대한 각 특징값을 기초로

$X_i = (X_{i,1}, L, X_{i,n})$ 에 대해 라벨 Y_i 를 할당하고, 로우 데이터 X_i 에 라벨 Y_i 를 매핑하여, $\{(X, Y)\} = \{(X_i, Y_i)\}_{i=1}^m$ 데이터 집합을 생성할 수 있다.

[0060] 이때 S1010 단계에서, 장치(100)는 로우 데이터를 신경망 모델에 적합한 형태로 가공하여 사용할 수 있도록, 로우 데이터를 전처리하는 동작을 수행할 수 있다.

[0061] 예를 들어, 장치(100)는 로우 데이터 집합 $\{(X, Y)\} = \{(X_i, Y_i)\}_{i=1}^m$ 에 대해 하기 수학식 1을 기초로 $\min$ - $\max$ 변환을 수행하여, 학습 집합 $\{(x_i, y_i)\}_{i=1}^{m_1}$ ($m_1 = m \times r$, r)은 상기 데이터 집합에서 학습 집

합에 사용될 비율, $r \in (0,1)$ 을 생성하는 동작을 포함할 수 있다.

수학적식 1

$$\min\text{-max}(x_{ij}) = \frac{x_{ij} - \min(\mathbf{x}^i)}{\max(\mathbf{x}^i) - \min(\mathbf{x}^i)}, j = 1, L, n,$$

($\mathbf{x}^i$ 는 $\{\mathbf{x}_i\}_{i=1}^m$ 의 i 번째 특징으로 이루어진 열벡터)

다음으로, S1020 단계에서, 장치(100)는 신경망의 은닉 레이어에 대한 정보를 설정할 수 있다. 장치(100)는 은닉 레이어의 활성화 함수 f , 은닉 레이어의 노드 개수 k_0 및 $\{k_p\}_{p=1}^P$ (p 는 은닉 레이어를 특정하는 인덱스)를 포함하는 은닉 레이어에 대한 정보를 설정할 수 있다.

예를 들어, 장치(100)는 일 실시 예에 따른 신경망을 ELM(Extreme Learning Machine) 또는 RVFL(Random Vector Functional Link)을 기초로 네트워크를 설계할 수 있다.

도 3은 일 실시 예에 따라 ELM(Extreme Learning Machine)을 기초로 설계한 신경망의 구조를 예시적으로 도시한 도면이다.

도 3을 참조하면, ELM은 딥러닝 신경망의 은닉층에서 사용되는 가중치 행렬 및 바이어스 등의 하이퍼파라미터를 조절할 필요가 없이, 여러 유형의 활성화 함수의 기능과 함께 작동할 수 있는 신경망 설계를 사용한다. 즉, ELM은 딥러닝 신경망에서 하이퍼파라미터의 튜닝 시간을 제거하여 빠르게 학습하므로 반복적인 실행으로 인한 시간 소모적인 기존 딥러닝의 단점을 극복한 것이며, 이러한 ELM의 설계에 본 발명의 실시 예를 적용하는 경우, 보다 효과적인 가중치를 초기 설정할 수 있다.

도 4는 일 실시 예에 따라 RVFL(Random Vector Functional Link)을 기초로 설계한 신경망의 구조를 예시적으로 도시한 도면이다.

도 4를 참조하면, RVFL은 하이퍼파라미터를 무작위로 선택하고 다양한 활성화 기능과 협력하여 빠르게 학습하는 신경망 설계를 사용한다. RVFL은 출력에 대한 가중치만 훈련하고 출력을 계산하기 위해 의사역행렬(pseudo-inverse)을 사용한다. 여기서, RVFL이 가지는 ELM과의 큰 차이점은 입력 데이터를 출력에 직접 링크하여 사용한다는 것이다. 이러한 RVFL의 설계에 본 발명의 실시 예를 적용하는 경우, 보다 효과적이 가중치를 초기 설정할 수 있다.

상술한 도 3 및 도 4의 신경망 설계는 이해의 편의를 돕기 위한 예시일 뿐, 본 발명의 실시 예가 반드시 위 예시에 한정되는 것은 아니다.

한편, S1020 단계의 활성화 함수 f , 은닉 레이어의 노드 개수 k_0 및 $\{k_p\}_{p=1}^P$ (p 는 은닉 레이어를 특정하는 인덱스)를 기초로 한 신경망의 정보는 아래와 같은 기호로 표현될 수 있다.

대문자 P 는 최종 은닉 레이어를 특정하는 인덱스이다. 소문자 p 는 각 은닉 레이어를 특정하는 인덱스이다.

$W^p = (w_{u,v}^p)$ 는 $(p-1)$ 번째 은닉 레이어에 있는 노드와 p 번째 은닉 레이어에 있는 노드 사이의 가

중치 행렬을 의미한다. $b^p = (b_u^p)$ 에서 b_u^p 는 P 번째 은닉 레이어에 있는 노드 u 에 할당된 바이어스 값을 의미한다.

[0073]

이때 가중치 행렬 $W^p = (w_{u,v}^p)$ 과 바이어스 행렬 $b^p = (b_u^p)$ 은 하기 수학식 2와 같이 나타낼 수 있다.

수학식 2

$$W^p = \begin{bmatrix} w_1^p & w_2^p & L & w_{k_p}^p \end{bmatrix}^T = \begin{bmatrix} w_{11}^p & w_{12}^p & L & w_{1k_p}^p \\ M & M & L & M \\ w_{k_p-1}^p & w_{k_p-2}^p & L & w_{k_p-k_p}^p \end{bmatrix}^T = \begin{bmatrix} w_{11}^p & w_{12}^p & L & w_{1k_p}^p \\ M & M & L & M \\ w_{1k_p}^p & w_{2k_p}^p & L & w_{k_p-k_p}^p \end{bmatrix}_{k_p \times k_p},$$

$$b^p = \begin{bmatrix} b_1^p \\ M \\ b_{k_p}^p \end{bmatrix},$$

[0074]

수학식 2에서 T 는 전치행렬을 나타내고, 활성화 함수 f 는 실수에서 정의된 유한한 함수값을 갖는 함수이다.

[0075]

수학식 2에 따르면 H_1 을 하기 수학식 3과 같이 표현할 수 있다.

수학식 3

$$H_1 = f(W^1 \cdot x + b^1) = \begin{bmatrix} f(w_1^1 \cdot x_1 + b_1^1) & f(w_2^1 \cdot x_1 + b_2^1) & L & f(w_k^1 \cdot x_1 + b_k^1) \\ M & M & L & M \\ f(w_1^1 \cdot x_m + b_1^1) & f(w_2^1 \cdot x_m + b_2^1) & L & f(w_k^1 \cdot x_m + b_k^1) \end{bmatrix}$$

$$= \begin{bmatrix} f(h_{11}^1 + b_1^1) & f(h_{12}^1 + b_2^1) & L & f(h_{1k}^1 + b_k^1) \\ M & M & L & M \\ f(h_{m1}^1 + b_1^1) & f(h_{m2}^1 + b_2^1) & L & f(h_{mk}^1 + b_k^1) \end{bmatrix}_{m \times k},$$

[0077]

H_1 를 편의상 $H_1 = W^1 x$ 로 나타내면 $\{H_p\}_{p=1}^P$ 는 하기 수학식 4와 같이 표현할 수 있다.

수학식 4

$$H_1 = W^1 x, H_2 = W^2 (W^1 x) = f(W^2 \cdot H_1 + b^2), L, H_p = W^p (W^{p-1} L W^1 x) = f(W^p \cdot H_{p-1} + b^p).$$

[0079]

ELM의 경우, p 번째 출력 행렬 H_p 는 하기 수학식 5와 같이 표현될 수 있다.

[0080]

수학식 5

$$\begin{aligned}
 H_p &= f(W^p \cdot H_{p-1} + b^p) \\
 &= \begin{bmatrix} f(w_1^p \cdot H_{p-1}(1) + b_1^p) & f(w_2^p \cdot H_{p-1}(1) + b_2^p) & L & f(w_{k_p}^p \cdot H_{p-1}(1) + b_{k_p}^p) \\ M & M & L & M \\ f(w_1^p \cdot H_{p-1}(k_{p-1}) + b_1^p) & f(w_2^p \cdot H_{p-1}(k_{p-1}) + b_2^p) & L & f(w_{k_p}^p \cdot H_{p-1}(k_{p-1}) + b_{k_p}^p) \end{bmatrix} \\
 &= \begin{bmatrix} f(h_{1,1}^p + b_1^p) & f(h_{1,2}^p + b_2^p) & L & f(h_{1,k_p}^p + b_{k_p}^p) \\ M & M & L & M \\ f(h_{k_{p-1},1}^p + b_1^p) & f(h_{k_{p-1},2}^p + b_2^p) & L & f(h_{k_{p-1},k_p}^p + b_{k_p}^p) \end{bmatrix}_{k_{p-1} \times k_p},
 \end{aligned}$$

수학식 5에서 $w_i^p \cdot H_{p-1}(j)$ 는 w_i^p 와 H_{p-1} 의 j 번째 행과의 내적이다. 최종 출력

y^{out} 는 출력벡터 $[y_1^{out} \ L \ y_m^{out}]^T$ 로 $H_p W^{out} = y^{out}$ 풀어서 구할 수 있다. 구체적으로, y_i^{out} 는 하기 수학식 6과 같이 구할 수 있다.

수학식 6

$$y_i^{out} = \sum_{j=1}^{k_p} w_j^{out} f[W^h \cdot H_{h-1}(i) + b_j^h] \equiv W^{out} \cdot H_h(i), i=1, L, m.$$

수학식 6에서 W^{out} 는 최종 은닉 레이어의 노드와 출력 레이어의 노드 사이의 가중치 행렬이다.

RVFL의 경우, 가중치 행렬, 바이어스 행렬은 수학식 2와 동일하나, 출력 행렬 H_p 는 하기 수학식 7과 같이 표현할 수 있다.

수학식 7

$$\begin{aligned}
 H_p &= \text{Concatenate} \left(f(W^p \cdot H_{p-1} + b^p), \begin{bmatrix} x_1 \\ M \\ x_m \end{bmatrix} \right) \\
 &= \begin{bmatrix} f(h_{1,1}^p + b_1^p) & f(h_{1,2}^p + b_2^p) & L & f(h_{1,m}^p + b_m^p) & x_{1,1} & L & x_{1,n} \\ M & M & L & M & M & L & M \\ f(h_{m,1}^p + b_1^p) & f(h_{m,2}^p + b_2^p) & L & f(h_{m,m}^p + b_m^p) & x_{m,1} & L & x_{m,n} \end{bmatrix}_{m \times (m+n)}.
 \end{aligned}$$

- [0087] 수학적식 7에서 k_{p-1} 와 k_p 는 모두 m이다.
- [0088] 이때, S1020 단계에서, 장치(100)는 상기 최종 은닉 레이어와 출력 레이어 사이의 가중치 행렬의 랭크 $r(W^p)$ 가 최대가 되도록, 은닉 레이어의 가중치 행렬 및 바이어스 행렬 $\{W^p, b^p\}_{p=1}^P$ 의 초기값을 설정할 수 있다.
- [0089] 예를 들어, 장치(100)는 입력 레이어로부터 출력 레이어로 향하는 방향으로 각 은닉 레이어에 대해, p번째(p는 은닉 레이어를 특징하는 인덱스) 레이어의 출력 함수 H_p 의 랭크 $r(H_p)$ 가 p번째 레이어의 출력 함수의 행의 개수 또는 열의 개수 중의 최소값($\min(k_{p-1}, k_p)$)과 같은지 판별할 수 있다. 이 경우, 장치(100)는 해당 판별에서 $r(H_p) = \min(k_{p-1}, k_p)$ 를 만족하지 않는다면, 기 설정된 횟수만큼 p번째 레이어의 바이어스를 갱신하여 p번째 레이어의 출력 함수 H_p 를 갱신함으로써, 가중치 행렬의 랭크 $r(W^p)$ 가 최대가 되도록 은닉 레이어의 가중치 행렬 및 바이어스 행렬 $\{W^p, b^p\}_{p=1}^P$ 의 초기값을 재설정할 수 있다.
- [0090] 다음으로, S1030 단계에서, 장치(100)는 학습 데이터를 이용해 얻은 예측값과 실제값 간의 오차 최소화를 기반으로 얻은 최종 은닉 레이어의 출력 행렬에 대한 랭크에 기초해 최종 가중치 행렬에 적용할 무어-펜로즈 일반화 역행렬을 결정하여 최종 가중치 행렬을 설정할 수 있다.
- [0091] S1030 단계의 구체적 예시를 설명하기에 앞서, 설명의 이해에 필요한 수학적 정의에 대해 먼저 설명하기로 한다.
- [0092] [정의 1]
- [0093] $HG = I_m$ ($GH = I_n$) 를 만족하는 경우, G 를 H 의 우(또는 좌)의사 역렬(right(left) pseudo-inverse)이라 정의한다. 여기서 I_m 는 $m \times m$ 단위행렬이다. 우의사 역행렬(right pseudo-inverse)과 좌의사 역행렬(left pseudo-inverse)은 각각 H^{r-} 와 H^{l-} 로 표기한다.
- [0094] [정의 2]
- [0095] $\text{rank}(H) = r(H) < \min(n, m)$ 라고 할 때,
- [0096] [정의 2]-(a).
- [0097] $HGH = H$ 및 $GHG = G$ 의 두 등식을 만족하면 G 를 H 의 일반화 역행렬(generalized inverse)이라 정의한다.
- [0098] [정의 2]-(b).
- [0099] $HGH = H$, $GHG = G$, $(HG)^T = HG$, $(GH)^T = GH$ 의 네 등식을 만족하면 G 를 H 의 무어-펜로즈 일반화 역행렬(Moore-Penrose generalized inverse)이라 정의한다. 이때 일반화 역행렬

(generalized inverse)은 H^- 로 표기하고, 무어-펜로즈 일반화 역행렬(Moore-Penrose generalized inverse)은 H^+ 로 표기하기로 한다.

이어서, S1030 단계에서 최종 가중치 행렬을 설정하는 구체적 예시를 설명한다.

먼저, 최종 가중치 행렬 W^{out} 은 $H_P W^{out} = y^{out}$ 에서 하기 수학식 8에 나타나 있는 H_P 의 무어-펜로즈 일반화 역행렬의 극한을 이용하여 구할 수 있다.

즉, 최종 가중치 행렬 W^{out} 은 $W^{out} = H_P^+ y^{out}$ 를 통해 도출할 수 있으며, 임의의 $m \times n$ 행렬 H 에 대하여 그 행렬의 M-P GI H^+ 는 다음과 같다.

수학식 8

$$H^+ = \lim_{\delta \rightarrow 0} (H^T H + \delta^2 I_n)^{-1} H^T = \lim_{\delta \rightarrow 0} H^T (H H^T + \delta^2 I_m)^{-1}.$$

여기서, $\text{rank}(H) = r(H) < \min(n, m)$ 라 하면, $G = H^+$ 가 될 수 있는 필요충분조건은

[정의 2]-(b)에 나타난 4개의 등식을 만족하는 것이다. 이때 H^+ 는 항상 존재하고 유일하게 존재한다.

이 경우, 최종 은닉 레이어에서 얻은 행렬 H_P 의 랭크에 따라 예측 오차 $E_2 = \|y^{out} - y\|_2$ (L_2 norm)를 확인할 수 있으며, 이는 하기 수학식 9와 같이 최종 레이어의 출력 행렬에 대한 랭크와 학습 데이터에 대한 오차값에 기초해 나타낼 수 있다.

즉, 실수에서 정의된 활성화 함수 f 가 임의의 유한값을 함수값으로 갖는다고 가정하고, $w_{i,j}^p$ 와 b_i^p 는 모든 $p=1, L, P$ 에 대해 각각 i 와 j 임의로 선택했다고 하면, 다음 수학식 9가 성립한다.

수학식 9

$$\left\{ \begin{array}{l} \text{(i) } r(H_P) = k_{p-1} \text{인 경우: } W^{out} = H_P^+ y^{out} \text{이고 } E_2 \equiv 0 \\ \text{(ii) } r(H_P) = k_p \text{인 경우} \\ \quad \text{(a) } y \in C(H_P): W^{out} = (H_P^T H_P)^{-1} H_P^T y^{out} \text{이고 } E_2 \equiv 0 \\ \quad \text{(b) } y \notin C(H_P): W^{apout} = (H_P^T H_P)^{-1} H_P^T y^{out} \text{일 때 } E_2 \text{는 최소값} \\ \text{(iii) } r(H_P) < \min(k_{p-1}, k_p) \text{인 경우} \\ \quad \text{(a) } y \in C(H_P): W^{out} \text{의 성분 중에는 자유변수가 있고 } E_2 \equiv 0 \\ \quad \text{(b) } y \notin C(H_P): W^{apout} = (H_P^T H_P)^+ H_P^T y^{out} \text{일 때 } E_2 \text{는 최소값} \end{array} \right.$$

[0108] 여기서, W^{apout} 는 오차가 0이 아닌 경우 가중치 행렬 W^{out} 을 근사한 근사 행렬이고, $C(H_P)$ 는 H_P 의 열공간(column space)이다. 또한, 수학식 9 (iii)의 (a)에서 W^{out} 의 성분 중 자유변수의 개수는 $\min(k_{P-1}, k_P) - r(H_P)$ 이다.

[0109] 수학식 9에 따르면, S1030 단계에서, 장치(100)는 최종 레이어의 출력 행렬 H_P 에 대한 랭크 $r(H_P)$ 에 따라 최종 가중치 행렬에 적용할 무어-펜로즈 일반화 역행렬 H_P^{r-} , $(H_P^T H_P)^{-1}$, $(H_P^T H_P)^+$ 을 결정하고, 하기 조건에 나타난 바와 같이 데이터의 예측값과 상기 데이터의 실제값의 L2 norm 오차 E2가 0 인지 또는 0보다 큰 실수인지를 구분한다. 이에 따라 상기 최종 가중치 행렬 W^{out} (E2가 0인 경우) 또는 W^{apout} (E2가 0이 아닌 경우)을 설정할 수 있다.

[0110] 다음으로, 실제 숫자를 사용한 예시를 통해, 상술한 도 2의 동작을 설명하기로 한다.

[0111] 먼저, S1010 단계에서, 학습 데이터 집합 $\{X_i\}_{i=1}^m$ 가 $\{(0,1), (1,0), (1,2)\}$ 로 주어졌다고 가정한다. 각 학습 데이터의 라벨은 0,1,1로 할당되었다고 한다. 라벨을 포함한 $\{(0,1), 0\}, \{(1,0), 1\}, \{(1,2), 1\}$ 의 집합을 이미 전처리된 데이터 집합으로 가정한다. 여기서, $m=3, n=2$ 이고 $m=k_1$ 이다.

[0112] 이후, S1020 단계에서, 은닉층이 2개이고 각 층이 3개의 노드로 구성된 네트워크를 가정한다. 여기서, $P=2, k_1=k_2=3$ 이다.

[0113] 또한, 초기 가중치 행렬 및 바이어스 행렬은 아래와 같이 주어졌다고 가정한다.

$$W^1 = [w_1^1 \ w_2^1 \ w_3^1]^T = \begin{bmatrix} 1/2 & 0 & 1 \\ 1/2 & 1 & 0 \end{bmatrix}^T = \begin{bmatrix} 1/2 & 1/2 \\ 0 & 1 \\ 1 & 0 \end{bmatrix},$$

[0114]

$$W^2 = [w_1^2 \ w_2^2 \ w_3^2]^T = \begin{bmatrix} w_1^2 \\ w_2^2 \\ w_3^2 \end{bmatrix} = \begin{bmatrix} 1/3 & 0 & 1 \\ 1/3 & 1 & 0 \\ 1/3 & 0 & 0 \end{bmatrix}^T = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix},$$

[0115]

$$b^1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix},$$

[0116]

$$\mathbf{b}^2 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

활성화 함수 f 는 아래의 ReLU 함수를 사용하는 것으로 설정한다.

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

이 경우, H_1 과 H_2 는 다음과 같이 계산될 수 있다.

$$H_1 = \begin{bmatrix} 1/2 & 0 & 1 \\ 1/2 & 1 & 0 \\ 3/2 & 2 & 1 \end{bmatrix}_{3 \times 3},$$

$$\begin{aligned} H_2 &= f(\mathbf{W}^2 \cdot H_1 + \mathbf{b}^{2T}) = f(H_1 \mathbf{W}^{2T} + \mathbf{b}^{2T}) = f \left(\begin{bmatrix} 1/2 & 0 & 1 \\ 1/2 & 1 & 0 \\ 3/2 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1/3 & 0 & 1 \\ 1/3 & 1 & 0 \\ 1/3 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \end{bmatrix} \right) \\ &= \begin{bmatrix} f(1/6+1/3+0) & f(0+0) & f(1/2+0) \\ f(1/6+1/3+0) & f(1+0) & f(1/2+0) \\ f(3/6+2/3+1/3+0) & f(2+0) & f(3/2+0) \end{bmatrix} \\ &= \begin{bmatrix} 1/2 & 0 & 1/2 \\ 1/2 & 1 & 1/2 \\ 3/2 & 2 & 3/2 \end{bmatrix}_{3 \times 3}. \end{aligned}$$

이 경우, S1030 단계의 동작을 계산하면, $\mathbf{y} \notin C(H_2)$ 임을 확인할 수 있다. 즉, 수학적 9에서 (iii)(b)에

해당한다. 이 경우, $H_2^T H_2$ 와, 그의 일반화 역행렬 $(H_2^T H_2)^-$ 는 다음과 같이 계산할 수 있다.

$$H_2^T H_2 = \begin{bmatrix} 11/4 & 7/2 & 11/4 \\ 7/2 & 5 & 7/2 \\ 11/4 & 7/2 & 11/4 \end{bmatrix}, \quad (H_2^T H_2)^- = \begin{bmatrix} 10/3 & -7/3 & 0 \\ -7/3 & 11/6 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

이에, 수학적 9에 따라, $\mathbf{W}^{apout}$ 는 다음과 같이 계산될 수 있다.

$$\mathbf{W}^{apout} = (H_2^T H_2)^- H_2^T \mathbf{y} = \begin{bmatrix} 10/3 & -7/3 & 0 \\ -7/3 & 11/6 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1/2 & 1/2 & 3/2 \\ 0 & 1 & 2 \\ 1/2 & 1/2 & 3/2 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 10/3 & -7/3 & 0 \\ -7/3 & 11/6 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1/2 \\ 0 \\ 1/2 \end{bmatrix} = \begin{bmatrix} 5/3 \\ -7/6 \\ 0 \end{bmatrix}.$$

또한, $\mathbf{y}^{out}$ 과 이에 대응하는 오차 E_2 는 각각 다음과 같다.

$$\mathbf{y}^{out} = \mathbf{W}^{apout} \cdot \mathbf{H}_2 = \begin{bmatrix} 1/2 & 0 & 1/2 \\ 1/2 & 1 & 1/2 \\ 3/2 & 2 & 3/2 \end{bmatrix} \begin{bmatrix} 5/3 \\ -7/6 \\ 0 \end{bmatrix} = \begin{bmatrix} 5/6 \\ -1/3 \\ 1/6 \end{bmatrix}$$

$$\|\mathbf{y}^{out} - \mathbf{y}\| = \sqrt{1/36 + 1/9 + 1/36} = 1/\sqrt{6} \approx 0.4082.$$

여기서, $(\mathbf{H}_2^T \mathbf{H}_2)^{-}$ 대신에 $(\mathbf{H}_2^T \mathbf{H}_2)^{+}$ 를 사용하면 오차는 더 작아질 수 있다. 즉, 만일 정리 1에서

$\delta = 10^{-6}$ 을 사용하여 M-P GI $(\mathbf{H}_2^T \mathbf{H}_2)^{+}$ 를 구하면 다음과 같다.

$$(\mathbf{H}_2^T \mathbf{H}_2)^{+} = \begin{bmatrix} 0.8324 & -1.1653 & 0.8324 \\ -1.1653 & 1.8313 & -1.1653 \\ 0.8324 & -1.1653 & 0.8324 \end{bmatrix}$$

이를 이용하면 다음과 같은 $\mathbf{W}^{apout}$ 과 $\mathbf{y}^{out}$ 을 얻을 수 있다.

$$\mathbf{W}^{apout} = [0.8324 \quad -1.1653 \quad 0.8324]^T, \quad \mathbf{y}^{out} = [0.8324 \quad -0.3329 \quad 0.1666]^T.$$

여기서, $\mathbf{y}^{out}$ 를 사용하여 구한 오차는 $E_2 \approx 0.166667$ 로 이론적으로 구할 수 있는 오차의 최소값에 해당한다.

이를 통해, 상술한 실시 예는 수학적 이론을 기반으로 학습 과정에서 파생된 행렬의 특성을 파악하고, 정확한 예측과 학습 속도에 중점을 둔 무어-펜로즈 일반화 역행렬을 이용한 극한 접근 방식을 적용함으로써, 가중치 학습에 소모되는 자원을 획기적으로 감소시키면서도 신경망 모델이 정확한 출력값을 도출하게 하여 예측 오차를 최소화할 수 있다.

다음으로, 도 5를 통해 상술한 도 2의 동작을 신경망의 정보를 표현하는 수학적식을 사용하여 설명하기로 한다.

도 5는 일 실시 예에 따른 신경망 생성 방법을 신경망의 정보를 표현하는 수학적식을 사용하여 예시적으로 도시한 도면이다. 도 5의 각 단계는 도 1의 장치(100)가 수행하는 것으로 이해될 수 있다. 도 5의 각 단계는 도 2의 동작을 알고리즘의 형태로 분할하여 예시한 것으로 이해될 수 있다.

도 5를 참조하면, STEP 1에서, 라벨이 생성된 데이터 집합 $\{(X, Y)\} = \{(X_i, Y_i)\}_{i=1}^m$ 을 획득할 수 있다.

STEP 2에서, 데이터 집합 $\{(X, Y)\} = \{(X_i, Y_i)\}_{i=1}^m$ 에 대해 전처리를 수행할 수 있으며, 전처리가 수행된

데이터를 $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^m$ 로 표현할 수 있다.

STEP 3에서, 전처리된 데이터 집합을 학습 집합(training set)과 테스트 집합(testing set)으로 나눌 수 있다.

신경망의 학습에 사용할 학습 집합은 $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{m_1}$, 신경망의 테스트에 사용할 테스트 집합은

$$\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=m_1+1}^m, \quad m_1 = m \times r,$$

로 표기할 수 있고, 여기서 r 은 전체 데이터 집합 중 학습 집합의 비율

$$r \in (0,1)$$

이다.

[0141]

STEP 4에서, 신경망의 활성화 함수 f , 은닉 레이어의 노드 개수 k_0 및 $\{k_p\}_{p=1}^P$ (p 는 은닉 레이어를 특정하는 인덱스)를 포함하는 네트워크 환경을 설정할 수 있다.

[0142]

STEP 5에서, 은닉 레이어의 가중치 행렬과 바이어스 행렬의 초기값을 설정할 수 있다. 이때 $\{W^p, b^p\}_{p=1}^P$ 설정은 상술한 정리 3을 고려하여 $r(W^p)$ 가 최대가 되도록 설정할 수 있다. 이와 같이 초기값을 설정함으로써 최적의 $\{W^p, b^p\}_{p=1}^P$ 를 구하기 위한 업데이트 (STEP 7~8) 횟수를 줄일 수 있다.

[0143]

이후, 신경망 모델의 학습 과정은 STEP 6~10에 해당한다.

[0144]

STEP 6에서, 초기 설정 $p=1$ (첫번째 은닉 레이어)부터 알고리즘을 시작한다.

[0145]

STEP 7-8을 통해, 입력 레이어로부터 출력 레이어로 향하는 방향으로 각 은닉 레이어에 대해, p 번째(p 는 은닉

레이어를 특정하는 인덱스) 레이어의 출력 함수 H_p 의 랭크 $r(H_p)$ 가 p 번째 레이어의 출력 함수의 행의 개수 또는 열의 개수 중의 최소값($\min(k_{p-1}, k_p)$)과 같은지 판별할 수 있다. 이 경우, 장치(100)는 해당 판별에서 $r(H_p) = \min(k_{p-1}, k_p)$ 를 만족하지 않는다면, 기 설정된 횟수만큼 p 번째 레이어의 바이어스를 갱신하여

p 번째 레이어의 출력 함수 H_p 를 갱신함으로써, 가중치 행렬의 랭크 $r(W^p)$ 가 최대가 되도록 은닉 레이어의 가중치 행렬 및 바이어스 행렬 $\{W^p, b^p\}_{p=1}^P$ 의 초기값을 재설정할 수 있다.

[0146]

STEP 7-8에서, $r(H_p) = \min(k_{p-1}, k_p)$ 인 경우 (a) $p < P$ 이면 STEP 8에서 p 를 1을 증가하고 다시 STEP 7로 이동한다. (b) $p = P$ 이면, $r(H_p)$ 에 따라 H_p^{T-} 나 $(H_p^T H_p)^{-1}$ 계산한다.

[0147]

STEP 7-8에서, $r(H_p) < \max(k_{p-1}, k_p)$ 인 경우 (a) b^p 가 업데이트가 가능하면 업데이트한 후 STEP 7로 이동한다. 이 경우 $b_i^p, i > r(H_p)$ 가 업데이트 된다. (b) 만일 b^p 를 기 설정한 횟수만큼 업데이트하였다면 $p = P$ 인지를 확인하고, 만일 $p < P$ 이면 p 를 1을 증가시키고 STEP 7로 이동한다.

[0148]

STEP 8에서, $p=P$ 라면, 은닉 레이어 중 최종 레이어의 출력 행렬에 대한 랭크와 학습 데이터에 대한 오차값에 기초해 최종 가중치 행렬에 적용할 무어-펜로즈 일반화 역행렬을 결정하여 신경망의 최종 가중치 행렬을 설정할

수 있다. 즉, H_p 을 계산하여, $r(H_p)$ 를 확인하고 그에 따라 $H_p^{r-}, (H_p^T H_p)^{-1}, (H_p^T H_p)^+$ 중 하나를 선택하여 계산한다.

[0149] STEP 9에서, 최종 가중치 행렬을 W^{out} 또는 W^{apout} 설정한다.

[0150] STEP 10에서, y_{tr}^{out} 와 E_2 계산한다. STEP 9에서 설정한 W^{out} 또는 W^{apout} 을 이용하여,

$y_{tr}^{r,out} = \{y_i^{r,out}\}_{i=1}^{m_1}$ 와 $y_{tr}^{l,out} = \{y_i^{l,out}\}_{i=1}^{m_1}$ 를 도출하고, 주어진 $\{y_i\}_{t=1}^{m_1}$ 와 비교하여 예측 오차 $E_{2,tr,r}$ 와 $E_{2,tr,l}$ 을 구할 수 있다. 여기서 $y_{tr}^{r,out}$ 와 $y_{tr}^{l,out}$ 은 각각 학습을 통해 얻은 출력으로 로우 데이터와 그 데이터를 라벨링한 데이터이고, $E_{2,tr,r}$ 와 $E_{2,tr,l}$ 는 각각 $y_{tr}^{r,out}$ 와 $y_{tr}^{l,out}$ 를 사용하여 얻은 오차이다.

[0151] 이후, 학습된 신경망 모델의 테스트 과정은 STEP 11-12에 해당한다.

[0152] STEP 11에서, W^{out} 및 W^{apout} 가 설정된 신경망에 테스트 집합 $\{(x_i, y_i)\}_{t=m_1+1}^m$ 에 적용하여

테스트 오차 $E_{2,te,r}$ 와 $E_{2,te,l}$ 를 계산한다. 여기서, 아랫 첨자 tr은 학습 데이터, te는 테스트 데이터, r 은 로우 데이터, l 은 라벨을 의미한다.

[0153] STEP 12에서, 만일 오차 $E_{2,te,r}$ 가 원하는 만큼 작으면 테스트 과정을 끝낸다. 그렇지 않으면 b^p 를 업데이트 한다. 이 경우 $b_i^p, i \leq r(H_p)$ 가 업데이트 될 수 있다.

[0154] 상술한 실시 예에 따르면, 수학적 이론을 기반으로 학습 과정에서 파생된 행렬의 특성을 파악하고, 정확한 예측과 학습 속도에 중점을 둔 무어-펜로즈 일반화 역행렬(Moore-Penrose generalized inverse)을 이용한 극한 접근 방식을 신경망 모델에 적용함으로써, 가중치 학습에 소모되는 자원을 획기적으로 감소시키면서도 정확한 출력을 구하여 예측 오차를 최소화할 수 있다.

[0155] 이에 따라, 시간과 자원이 많이 소요되는 기존의 가중치 학습 방법에 대한 단점을 극복하고, 본 발명은 학습에 적은 시간이 소요되면서도 높은 정확도의 예측이 가능할 뿐만 아니라 기존의 기술들에 비해 적은 자원을 사용하여 친환경적이다.

[0156] 또한, 상술한 실시 예에 따른 방법은 컴퓨터로 판독 가능한 매체에 저장된 컴퓨터 프로그램 또는 컴퓨터로 판독 가능한 기록 매체의 형태로도 구현될 수 있다. 일 실시 예에 따른 컴퓨터 판독 가능 기록매체에 저장되어 있는 컴퓨터 프로그램은 프로세서에 의해 실행되면, 하나 이상의 프로세서(10)에 의해, 신경망의 학습에 사용할 학습 데이터를 획득하는 동작; 신경망의 은닉 레이어에 대한 정보를 설정하는 동작; 및 은닉 레이어 중 최종 레이어의 출력 행렬에 대한 랭크와 학습 데이터에 대한 오차값에 기초해 최종 가중치 행렬에 적용할 무어-펜로즈 일반화 역행렬을 결정하여 최종 가중치 행렬을 설정하는 동작을 포함하는 방법을 프로세서가 수행하도록 하기 위한 명령어를 포함할 수 있다.

[0157] 또한, 본 문서의 실시 예들의 프로세서는, 저장 매체로부터 저장된 하나 이상의 명령어들 중 적어도 하나의 명령을 호출하고, 그것을 실행할 수 있다. 이것은 기기가 호출된 적어도 하나의 명령어에 따라 적어도 하나의 기능을 수행하도록 운영되는 것을 가능하게 한다. 이러한 하나 이상의 명령어들은 컴파일러에 의해 생성된 코드 또는 인터프리터에 의해 실행될 수 있는 코드를 포함할 수 있다. 프로세서는 범용 프로세서, FPGA(Field Programmable Gate Array), ASIC(Application Specific Integrated Circuit), DSP(Digital Signal Processor) 및/또는 등등 일 수 있다.

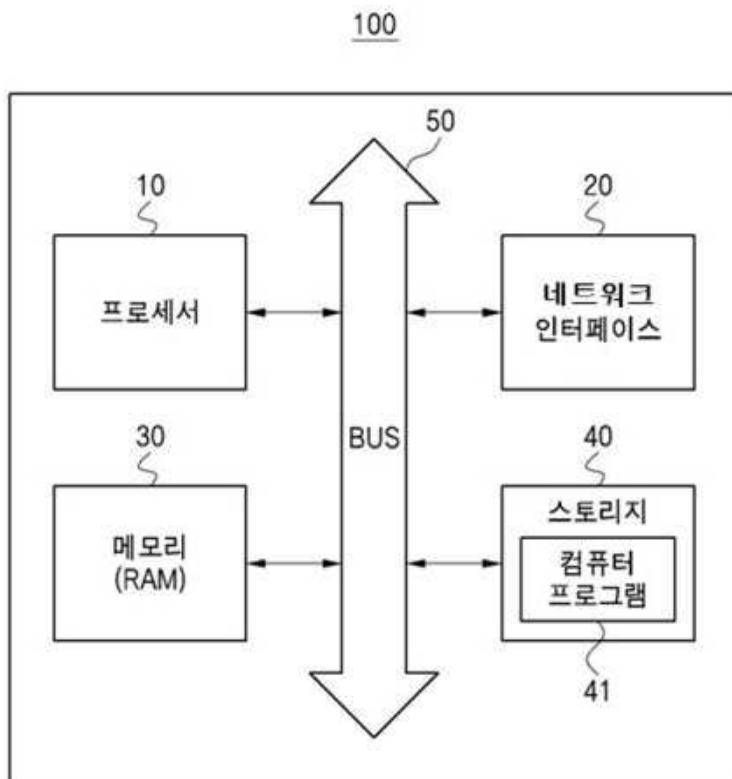
[0158] 이상 첨부된 도면을 참조하여 본 발명의 실시 예들을 설명하였지만, 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자는 본 발명이 그 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 실시될 수 있다는 것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시 예들은 모든 면에서 예시적인 것이며 한정적이 아닌 것으로 이해해야만 한다.

부호의 설명

- [0160]
- 10: 프로세서
 - 20: 네트워크 인터페이스
 - 30: 메모리
 - 40: 스토리지
 - 41: 컴퓨터 프로그램
 - 50: 데이터 버스
 - 100: 신경망 생성 장치

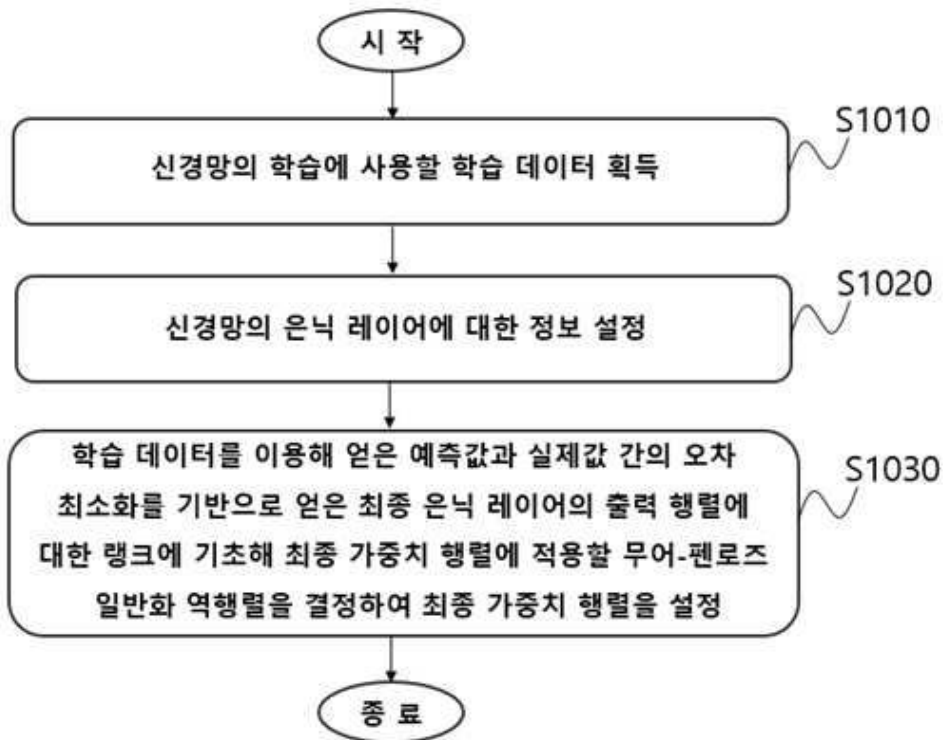
도면

도면1

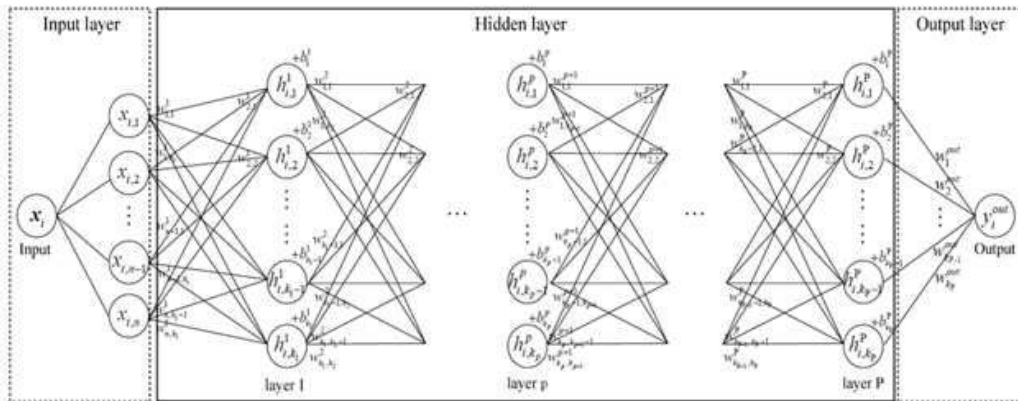


20

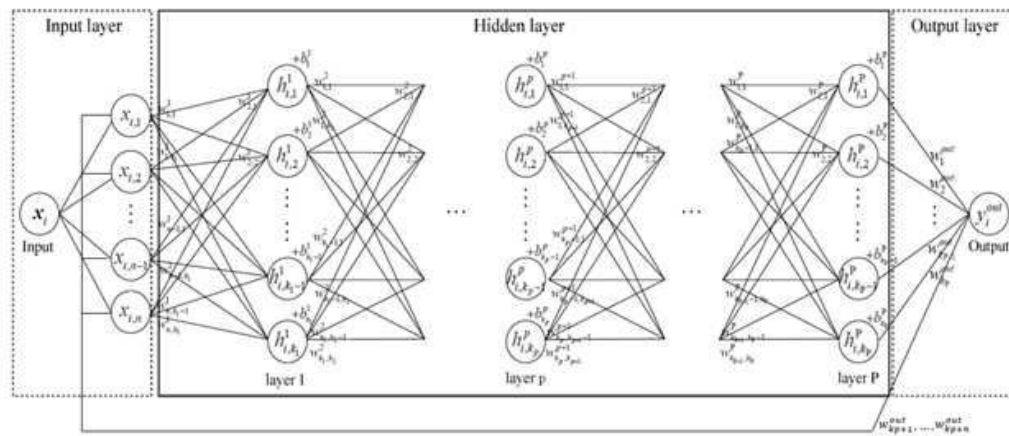
도면2



도면3



도면4



도면5

