



(19) 대한민국특허청(KR)

(12) 등록특허공보(B1)

(45) 공고일자 2023년09월26일

(11) 등록번호 10-2583938

(24) 등록일자 2023년09월22일

(51) 국제특허분류(Int. Cl.)

G06N 3/063 (2023.01) G06N 3/04 (2023.01)

G06N 3/08 (2023.01)

(52) CPC특허분류

G06N 3/063 (2013.01)

G06N 3/049 (2023.01)

(21) 출원번호 10-2021-0108887

(22) 출원일자 2021년08월18일

심사청구일자 2021년08월18일

(65) 공개번호 10-2023-0026813

(43) 공개일자 2023년02월27일

(56) 선행기술조사문헌

KR1020210061800 A

(뒷면에 계속)

(73) 특허권자

연세대학교 산학협력단

서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)

(72) 발명자

노원우

서울특별시 강남구 삼성로51길 35, 201동 1202호 (대치동, 래미안 대치 팰리스)

김민규

서울특별시 서대문구 연희로8길 28, 301호(연희동)

박천준

서울특별시 서대문구 연세로5다길 22-8, 204호(창천동)

(74) 대리인

특허법인우인

전체 청구항 수 : 총 6 항

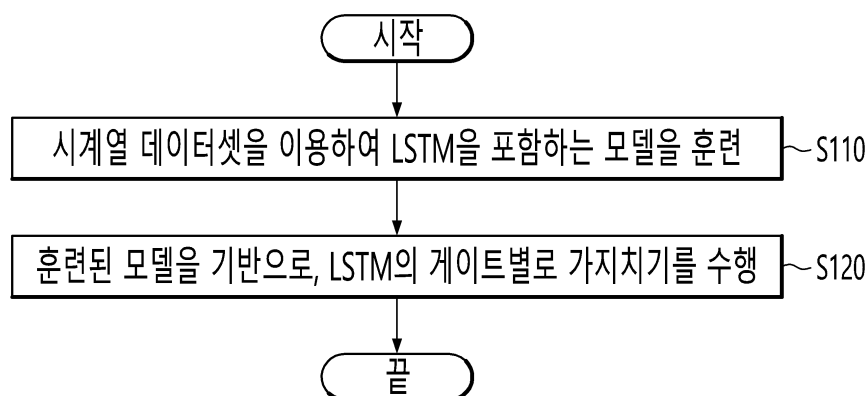
심사관 : 이준상

(54) 발명의 명칭 LSTM 가속을 위한 게이트-단위 가지치기 방법 및 장치

## (57) 요약

본 발명의 바람직한 실시예에 따른 LSTM 가속을 위한 게이트-단위 가지치기 방법 및 장치는, LSTM(Long Short-Term Memory)의 가지치기(pruning)를 게이트(gate) 단위로 수행함으로써 메모리 사용량을 줄일 수 있고, 게이트별로 동일한 가지치기 비율(pruning ratio)을 이용하여 게이트 단위로 LSTM의 가지치기를 수행함으로써 게이트 간 가지치기 비율의 불균형으로 인한 시계열 데이터의 처리 속도 감소를 해결할 수 있다.

대표도 - 도2



(52) CPC특허분류  
G06N 3/082 (2023.01)

(56) 선행기술조사문헌  
KR102256288 B1  
KR102256289 B1  
KR1020210012882 A  
KR1020200098121 A

이 발명을 지원한 국가연구개발사업

과제고유번호	1711133823
과제번호	10080674
부처명	산업통상자원부
과제관리(전문)기관명	한국산업기술평가관리원
연구사업명	산업기술혁신사업
연구과제명	재구성 가능한 인공신경망 가속기 구현 및 인스트럭션셋 기술개발(2/2)
기 여 율	1/1
과제수행기관명	연세대학교 산학협력단
연구기간	2021.01.01 ~ 2021.12.31

---

## 명세서

### 청구범위

#### 청구항 1

LSTM(Long Short-Term Memory)의 가지치기(pruning)를 수행하는 게이트-단위 가지치기 장치가 수행하는 게이트-단위 가지치기 방법으로서,

시계열 데이터셋(dataset)을 이용하여 상기 LSTM을 포함하는 모델을 훈련하는 단계; 및

훈련된 상기 모델을 기반으로, 상기 LSTM의 망각 게이트(Forget Gate), 상기 LSTM의 입력 게이트(Input Gate), 상기 LSTM의 입력 업데이트 게이트(Input Update Gate) 및 상기 LSTM의 출력 게이트(Output Gate) 각각에 대한 가지치기를 수행하는 단계;

를 포함하고,

상기 가지치기 수행 단계는,

미리 설정된 가지치기 비율(pruning ratio)을 기반으로, 상기 망각 게이트, 상기 입력 게이트, 상기 입력 업데이트 게이트 및 상기 출력 게이트 각각에 대한 가지치기를 수행하는 것으로 이루어지고,

상기 가지치기 수행 단계는,

상기 망각 게이트, 상기 입력 게이트, 상기 입력 업데이트 게이트 및 상기 출력 게이트 각각에 대해, 입력의 벡터에 대한 행렬-벡터 곱 연산과 히든 상태의 벡터에 대한 행렬-벡터 곱 연산의 2개의 행렬-벡터 곱 연산을 통합(concatenate)하여 하나의 행렬-벡터 곱 연산에 대한 하나의 가중치 행렬을 획득하고,

상기 망각 게이트, 상기 입력 게이트, 상기 입력 업데이트 게이트 및 상기 출력 게이트 각각에 대해 획득한 상기 하나의 가중치 행렬을 기반으로 가지치기를 수행하는 것으로 이루어지며,

상기 가지치기 수행 단계는,

상기 하나의 가중치 행렬을 기반으로, 열 벡터(column vector) 단위의 가지치기와 행 벡터(row vector) 단위의 가지치기를 수행하는 것으로 이루어지는,

LSTM 가속을 위한 게이트-단위 가지치기 방법.

#### 청구항 2

삭제

#### 청구항 3

제1항에서,

상기 가지치기 수행 단계는,

상기 망각 게이트, 상기 입력 게이트, 상기 입력 업데이트 게이트 및 상기 출력 게이트 각각에 대한 가지치기를 서로 동일한 상기 가지치기 비율을 기반으로 수행하는 것으로 이루어지는,

LSTM 가속을 위한 게이트-단위 가지치기 방법.

#### 청구항 4

삭제

#### 청구항 5

삭제

#### 청구항 6

제1항에서,

상기 가지치기 수행 단계는,

상기 하나의 가중치 행렬을 기반으로, 각 열 벡터(column vector)의 절대값 평균을 획득하고, 상기 가지치기 비율이 되도록 절대값 평균이 작은 순서대로 열 벡터를 제거하여 열 벡터 단위의 가지치기를 수행하며,

상기 하나의 가중치 행렬을 기반으로, 각 행 벡터(row vector)의 절대값 평균을 획득하고, 상기 가지치기 비율이 되도록 절대값 평균이 작은 순서대로 행 벡터를 제거하여 행 벡터 단위의 가지치기를 수행하는 것으로 이루어지는,

LSTM 가속을 위한 게이트-단위 가지치기 방법.

#### 청구항 7

제1항, 제3항, 및 제6항 중 어느 한 항에 기재된 LSTM 가속을 위한 게이트-단위 가지치기 방법을 컴퓨터에서 실행시키기 위하여 컴퓨터 판독 가능한 저장 매체에 저장된 컴퓨터 프로그램.

#### 청구항 8

LSTM(Long Short-Term Memory)의 가지치기(pruning)를 수행하는 게이트-단위 가지치기 장치로서,

상기 LSTM의 가지치기를 수행하기 위한 하나 이상의 프로그램을 저장하는 메모리; 및

상기 메모리에 저장된 상기 하나 이상의 프로그램에 따라 상기 LSTM의 가지치기를 수행하기 위한 동작을 수행하는 하나 이상의 프로세서;

를 포함하고,

상기 프로세서는,

시계열 데이터셋(dataset)을 이용하여 상기 LSTM을 포함하는 모델을 훈련하고,

훈련된 상기 모델을 기반으로, 상기 LSTM의 망각 게이트(Forget Gate), 상기 LSTM의 입력 게이트(Input Gate), 상기 LSTM의 입력 업데이트 게이트(Input Update Gate) 및 상기 LSTM의 출력 게이트(Output Gate) 각각에 대한 가지치기를 수행하고,

상기 프로세서는,

미리 설정된 가지치기 비율(pruning ratio)을 기반으로, 상기 망각 게이트, 상기 입력 게이트, 상기 입력 업데이트 게이트 및 상기 출력 게이트 각각에 대한 가지치기를 수행하고,

상기 프로세서는,

상기 망각 게이트, 상기 입력 게이트, 상기 입력 업데이트 게이트 및 상기 출력 게이트 각각에 대해, 입력의 벡터에 대한 행렬-벡터 곱 연산과 히든 상태의 벡터에 대한 행렬-벡터 곱 연산의 2개의 행렬-벡터 곱 연산을 통합(concatenate)하여 하나의 행렬-벡터 곱 연산에 대한 하나의 가중치 행렬을 획득하고,

상기 망각 게이트, 상기 입력 게이트, 상기 입력 업데이트 게이트 및 상기 출력 게이트 각각에 대해 획득한 상기 하나의 가중치 행렬을 기반으로 가지치기를 수행하고,

상기 프로세서는,

상기 하나의 가중치 행렬을 기반으로, 열 벡터(column vector) 단위의 가지치기와 행 벡터(row vector) 단위의 가지치기를 수행하는,

LSTM 가속을 위한 게이트-단위 가지치기 장치.

#### 청구항 9

삭제

#### 청구항 10

삭제

## 청구항 11

삭제

## 청구항 12

제8항에서,

상기 프로세서는,

상기 하나의 가중치 행렬을 기반으로, 각 열 벡터(column vector)의 절대값 평균을 획득하고, 상기 가지치기 비율이 되도록 절대값 평균이 작은 순서대로 열 벡터를 제거하여 열 벡터 단위의 가지치기를 수행하며,

상기 하나의 가중치 행렬을 기반으로, 각 행 벡터(row vector)의 절대값 평균을 획득하고, 상기 가지치기 비율이 되도록 절대값 평균이 작은 순서대로 행 벡터를 제거하여 행 벡터 단위의 가지치기를 수행하는,

LSTM 가속을 위한 게이트-단위 가지치기 장치.

## 발명의 설명

### 기술 분야

[0001] 본 발명은 LSTM 가속을 위한 게이트-단위 가지치기 방법 및 장치에 관한 것으로서, 더욱 상세하게는 LSTM(Long Short-Term Memory)의 가지치기(pruning)를 수행하는, 방법 및 장치에 관한 것이다.

### 배경 기술

[0002] LSTM(Long Short-Term Memory)은 음성 인식, 필기 인식, 시계열 예측, 동작 인식 등 많은 영역에서 사용되고 있다. LSTM의 구조 상 여러 셀(cell)을 쌓을수록 가중치를 저장하기 위해 필요한 메모리 용량이 증가한다. 또한, 최근에는 더 많은 정보와 연산량을 다루게 됨에 따라 모델의 크기가 커져 LSTM의 파라미터(parameter) 수가 늘어나고 연산 시간도 길어지고 있다.

[0003] 파라미터의 수를 줄이기 위해 비구조화 가지치기(unstructured pruning)이 처음 제안되었지만, 불규칙적인 희소 행렬(sparse matrix)의 곱 연산을 위해서는 추가적인 인덱싱(indexing)이 필요하다. 인덱싱 오버헤드(indexing overhead)를 없애기 위해 엘리먼트(element) 단위의 가지치기(pruning)가 아닌 가중치 행렬 상에서 열(column)이나 행(row) 벡터 단위로 제거하는 구조화 가지치기(structured pruning)가 제안되었다. 하지만, 구조화 가지치기는 제거하는 벡터의 크기가 크기 때문에 데이터 손실(data loss)이 발생하고 가지치기 비율(pruning ratio)을 높이는 데에 어려움이 있다.

[0004] DNN(Deep Neural Network)에 대한 가지치기 연구는 진행되고 있고, LSTM에도 적용가능하지만 LSTM내의 게이트마다 행렬-벡터 곱 연산이 지배적인 LSTM의 특징에 최적화되어 있지 않다.

## 발명의 내용

### 해결하려는 과제

[0005] 본 발명이 이루고자 하는 목적은, LSTM(Long Short-Term Memory)의 가지치기(pruning)를 게이트(gate) 단위로 수행하는, LSTM 가속을 위한 게이트-단위 가지치기 방법 및 장치를 제공하는 데 있다.

[0006] 또한, 본 발명이 이루고자 하는 목적은, 게이트별로 동일한 가지치기 비율(pruning ratio)을 이용하여 게이트 단위로 LSTM의 가지치기를 수행하는, LSTM 가속을 위한 게이트-단위 가지치기 방법 및 장치를 제공하는 데 있다.

[0007] 본 발명의 명시되지 않은 또 다른 목적들은 하기의 상세한 설명 및 그 효과로부터 용이하게 추론할 수 있는 범위 내에서 추가적으로 고려될 수 있다.

### 과제의 해결 수단

[0008] 상기의 목적을 달성하기 위한 본 발명의 바람직한 실시예에 따른 LSTM 가속을 위한 게이트-단위 가지치기 방법

은, LSTM(Long Short-Term Memory)의 가지치기(pruning)를 수행하는 게이트-단위 가지치기 장치가 수행하는 게이트-단위 가지치기 방법으로서, 시계열 데이터셋(dataset)을 이용하여 상기 LSTM을 포함하는 모델을 훈련하는 단계; 및 훈련된 상기 모델을 기반으로, 상기 LSTM의 망각 게이트(Forget Gate), 상기 LSTM의 입력 게이트(Input Gate), 상기 LSTM의 입력 업데이트 게이트(Input Update Gate) 및 상기 LSTM의 출력 게이트(Output Gate) 각각에 대한 가지치기를 수행하는 단계;를 포함하다.

- [0009] 여기서, 상기 가지치기 수행 단계는, 미리 설정된 가지치기 비율(pruning ratio)을 기반으로, 상기 망각 게이트, 상기 입력 게이트, 상기 입력 업데이트 게이트 및 상기 출력 게이트 각각에 대한 가지치기를 수행하는 것으로 이루어질 수 있다.
- [0010] 여기서, 상기 가지치기 수행 단계는, 상기 망각 게이트, 상기 입력 게이트, 상기 입력 업데이트 게이트 및 상기 출력 게이트 각각에 대한 가지치기를 서로 동일한 상기 가지치기 비율을 기반으로 수행하는 것으로 이루어질 수 있다.
- [0011] 여기서, 상기 가지치기 수행 단계는, 상기 망각 게이트, 상기 입력 게이트, 상기 입력 업데이트 게이트 및 상기 출력 게이트 각각에 대해, 게이트에 대한 복수개의 행렬-벡터 곱 연산을 통합(concatenate)하여 하나의 행렬-벡터 곱 연산에 대한 하나의 가중치 행렬을 획득하고, 상기 망각 게이트, 상기 입력 게이트, 상기 입력 업데이트 게이트 및 상기 출력 게이트 각각에 대해 획득한 상기 하나의 가중치 행렬을 기반으로 가지치기를 수행하는 것으로 이루어질 수 있다.
- [0012] 여기서, 상기 가지치기 수행 단계는, 상기 하나의 가중치 행렬을 기반으로, 열(column) 단위의 가지치기와 행(row) 단위의 가지치기를 수행하는 것으로 이루어질 수 있다.
- [0013] 여기서, 상기 가지치기 수행 단계는, 상기 하나의 가중치 행렬을 기반으로, 각 열 벡터(column vector)의 절대값 평균을 획득하고, 상기 가지치기 비율이 되도록 절대값 평균이 작은 순서대로 열 벡터를 제거하여 열 단위의 가지치기를 수행하며, 상기 하나의 가중치 행렬을 기반으로, 각 행 벡터(row vector)의 절대값 평균을 획득하고, 상기 가지치기 비율이 되도록 절대값 평균이 작은 순서대로 행 벡터를 제거하여 행 단위의 가지치기를 수행하는 것으로 이루어질 수 있다.
- [0015] 상기의 기술적 과제를 달성하기 위한 본 발명의 바람직한 실시예에 따른 컴퓨터 프로그램은 컴퓨터 판독 가능한 저장 매체에 저장되어 상기한 LSTM 가속을 위한 게이트-단위 가지치기 방법 중 어느 하나를 컴퓨터에서 실행시킨다.
- [0017] 상기의 목적을 달성하기 위한 본 발명의 바람직한 실시예에 따른 LSTM 가속을 위한 게이트-단위 가지치기 장치는, LSTM(Long Short-Term Memory)의 가지치기(pruning)를 수행하는 게이트-단위 가지치기 장치로서, 상기 LSTM의 가지치기를 수행하기 위한 하나 이상의 프로그램을 저장하는 메모리; 및 상기 메모리에 저장된 상기 하나 이상의 프로그램에 따라 상기 LSTM의 가지치기를 수행하기 위한 동작을 수행하는 하나 이상의 프로세서;를 포함하고, 상기 프로세서는, 시계열 데이터셋(dataset)을 이용하여 상기 LSTM을 포함하는 모델을 훈련하고, 훈련된 상기 모델을 기반으로, 상기 LSTM의 망각 게이트(Forget Gate), 상기 LSTM의 입력 게이트(Input Gate), 상기 LSTM의 입력 업데이트 게이트(Input Update Gate) 및 상기 LSTM의 출력 게이트(Output Gate) 각각에 대한 가지치기를 수행한다.
- [0018] 여기서, 상기 프로세서는, 미리 설정된 가지치기 비율(pruning ratio)을 기반으로, 상기 망각 게이트, 상기 입력 게이트, 상기 입력 업데이트 게이트 및 상기 출력 게이트 각각에 대한 가지치기를 수행할 수 있다.
- [0019] 여기서, 상기 프로세서는, 상기 망각 게이트, 상기 입력 게이트, 상기 입력 업데이트 게이트 및 상기 출력 게이트 각각에 대해, 게이트에 대한 복수개의 행렬-벡터 곱 연산을 통합(concatenate)하여 하나의 행렬-벡터 곱 연산에 대한 하나의 가중치 행렬을 획득하고, 상기 망각 게이트, 상기 입력 게이트, 상기 입력 업데이트 게이트 및 상기 출력 게이트 각각에 대해 획득한 상기 하나의 가중치 행렬을 기반으로 가지치기를 수행할 수 있다.
- [0020] 여기서, 상기 프로세서는, 상기 하나의 가중치 행렬을 기반으로, 열(column) 단위의 가지치기와 행(row) 단위의 가지치기를 수행할 수 있다.
- [0021] 여기서, 상기 프로세서는, 상기 하나의 가중치 행렬을 기반으로, 각 열 벡터(column vector)의 절대값 평균을 획득하고, 상기 가지치기 비율이 되도록 절대값 평균이 작은 순서대로 열 벡터를 제거하여 열 단위의 가지치기를 수행하며, 상기 하나의 가중치 행렬을 기반으로, 각 행 벡터(row vector)의 절대값 평균을 획득하고, 상기 가지치기 비율이 되도록 절대값 평균이 작은 순서대로 행 벡터를 제거하여 행 단위의 가지치기를 수행할 수 있다.

다.

### 발명의 효과

- [0022] 본 발명의 바람직한 실시예에 따른 LSTM 가속을 위한 게이트-단위 가지치기 방법 및 장치에 의하면, LSTM(Long Short-Term Memory)의 가지치기(pruning)를 게이트(gate) 단위로 수행함으로써, 메모리 사용량을 줄일 수 있다.
- [0023] 또한, 본 발명은 게이트별로 동일한 가지치기 비율(pruning ratio)을 이용하여 게이트 단위로 LSTM의 가지치기를 수행함으로써, 게이트 간 가지치기 비율의 불균형으로 인한 시계열 데이터의 처리 속도 감소를 해결할 수 있다.
- [0024] 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

- [0025] 도 1은 본 발명의 바람직한 실시예에 따른 LSTM 가속을 위한 게이트-단위 가지치기 장치를 설명하기 위한 블록도이다.
- 도 2는 본 발명의 바람직한 실시예에 따른 LSTM 가속을 위한 게이트-단위 가지치기 방법을 설명하기 위한 흐름도이다.
- 도 3은 본 발명의 바람직한 실시예에 따른 LSTM의 게이트를 설명하기 위한 도면이다.
- 도 4는 도 2에 도시한 가지치기 수행 단계를 설명하기 위한 흐름도이다.
- 도 5는 본 발명의 바람직한 실시예에 따른 LSTM의 게이트별 가중치 행렬을 설명하기 위한 도면이다.
- 도 6은 본 발명의 바람직한 실시예에 따른 열 단위 가지치기를 설명하기 위한 도면이다.
- 도 7은 본 발명의 바람직한 실시예에 따른 LSTM의 내부 연산 순서를 설명하기 위한 도면이다.

### 발명을 실시하기 위한 구체적인 내용

- [0026] 이하, 첨부된 도면을 참조하여 본 발명의 실시예를 상세히 설명한다. 본 발명의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 상세하게 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나, 본 발명은 이하에서 개시되는 실시예들에 한정되는 것이 아니라 서로 다른 다양한 형태로 구현될 수 있으며, 단지 본 실시예들은 본 발명의 개시가 완전하도록 하고, 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에게 발명의 범주를 완전하게 알려주기 위해 제공되는 것이며, 본 발명은 청구항의 범주에 의해 정의될 뿐이다. 명세서 전체에 걸쳐 동일 참조 부호는 동일 구성 요소를 지칭한다.
- [0027] 다른 정의가 없다면, 본 명세서에서 사용되는 모든 용어(기술 및 과학적 용어를 포함)는 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에게 공통적으로 이해될 수 있는 의미로 사용될 수 있을 것이다. 또한, 일반적으로 사용되는 사전에 정의되어 있는 용어들은 명백하게 특별히 정의되어 있지 않는 한 이상적으로 또는 과도하게 해석되지 않는다.
- [0028] 본 명세서에서 "제1", "제2" 등의 용어는 하나의 구성 요소를 다른 구성 요소로부터 구별하기 위한 것으로, 이들 용어들에 의해 권리범위가 한정되어서는 아니 된다. 예컨대, 제1 구성 요소는 제2 구성 요소로 명명될 수 있고, 유사하게 제2 구성 요소도 제1 구성 요소로 명명될 수 있다.
- [0029] 본 명세서에서 각 단계들에 있어 식별부호(예컨대, a, b, c 등)는 설명의 편의를 위하여 사용되는 것으로 식별 부호는 각 단계들의 순서를 설명하는 것이 아니며, 각 단계들은 문맥상 명백하게 특정 순서를 기재하지 않는 이상 명기된 순서와 다르게 일어날 수 있다. 즉, 각 단계들은 명기된 순서와 동일하게 일어날 수도 있고 실질적으로 동시에 수행될 수도 있으며 반대의 순서대로 수행될 수도 있다.
- [0030] 본 명세서에서, "가진다", "가질 수 있다", "포함한다" 또는 "포함할 수 있다" 등의 표현은 해당 특징(예컨대, 수치, 기능, 동작, 또는 부품 등의 구성 요소)의 존재를 가리키며, 추가적인 특징의 존재를 배제하지 않는다.
- [0033] 이하에서 첨부한 도면을 참조하여 본 발명에 따른 LSTM 가속을 위한 게이트-단위 가지치기 방법 및 장치의 바람직한 실시예에 대해 상세하게 설명한다.



- [0035] 먼저, 도 1을 참조하여 본 발명의 바람직한 실시예에 따른 LSTM 가속을 위한 게이트-단위 가지치기 장치에 대하여 설명한다.
- [0036] 도 1은 본 발명의 바람직한 실시예에 따른 LSTM 가속을 위한 게이트-단위 가지치기 장치를 설명하기 위한 블록도이다.
- [0037] 도 1을 참조하면, 본 발명의 바람직한 실시예에 따른 LSTM 가속을 위한 게이트-단위 가지치기 장치(이하 '게이트-단위 가지치기 장치'라 한다)(100)는 LSTM(Long Short-Term Memory)의 가지치기(pruning)를 게이트(gate)단위로 수행할 수 있다. 이로 인해, 본 발명은 메모리 사용량을 줄일 수 있다.
- [0038] 또한, 게이트-단위 가지치기 장치(100)는 게이트별로 동일한 가지치기 비율(pruning ratio)을 이용하여 게이트단위로 LSTM의 가지치기를 수행할 수 있다. 이로 인해, 본 발명은 게이트 간 가지치기 비율의 불균형으로 인한 시계열 데이터의 처리 속도 감소를 해결할 수 있다.
- [0039] 즉, 종래의 구조화 가지치기(structured pruning)는 행렬-벡터 곱 연산이 지배적인 LSTM의 특징에 최적화되어 있지 않아, LSTM에는 적용하기 어려운 문제가 있다. LSTM은 망각 게이트(Forget Gate), 입력 게이트(Input Gate), 입력 업데이트 게이트(Input Update Gate) 및 출력 게이트(Output Gate)와 같은 4개의 게이트를 포함하고 있고, 각각의 게이트에는 히든 상태(hidden state)의 벡터와 입력(input)의 벡터가 사용되며, 게이트 마다의 가중치를 가지고 있다. 이에, 본 발명은 이와 같은 LSTM의 특징을 고려하여, 가지치기를 게이트 단위로 수행할 수 있다. 또한, 본 발명은 게이트별로 동일한 가지치기 비율을 이용하여 게이트 단위로 LSTM의 가지치기를 수행할 수 있다.
- [0041] 이를 위해, 게이트-단위 가지치기 장치(100)는 하나 이상의 프로세서(110), 컴퓨터 판독 가능한 저장 매체(130) 및 통신 버스(150)를 포함할 수 있다.
- [0042] 프로세서(110)는 게이트-단위 가지치기 장치(100)가 동작하도록 제어할 수 있다. 예컨대, 프로세서(110)는 컴퓨터 판독 가능한 저장 매체(130)에 저장된 하나 이상의 프로그램(131)을 실행할 수 있다. 하나 이상의 프로그램(131)은 하나 이상의 컴퓨터 실행 가능 명령어를 포함할 수 있으며, 컴퓨터 실행 가능 명령어는 프로세서(110)에 의해 실행되는 경우 게이트-단위 가지치기 장치(100)로 하여금 LSTM의 가지치기(pruning)를 수행하기 위한 동작을 수행하도록 구성될 수 있다.
- [0043] 컴퓨터 판독 가능한 저장 매체(130)는 LSTM의 가지치기(pruning)를 수행하기 위한 컴퓨터 실행 가능 명령어 내지 프로그램 코드, 프로그램 데이터 및/또는 다른 적합한 형태의 정보를 저장하도록 구성된다. 컴퓨터 판독 가능한 저장 매체(130)에 저장된 프로그램(131)은 프로세서(110)에 의해 실행 가능한 명령어의 집합을 포함한다. 일 실시예에서, 컴퓨터 판독 가능한 저장 매체(130)는 메모리(랜덤 액세스 메모리와 같은 휘발성 메모리, 비휘발성 메모리, 또는 이들의 적절한 조합), 하나 이상의 자기 디스크 저장 디바이스들, 광학 디스크 저장 디바이스들, 플래시 메모리 디바이스들, 그 밖에 게이트-단위 가지치기 장치(100)에 의해 액세스되고 원하는 정보를 저장할 수 있는 다른 형태의 저장 매체, 또는 이들의 적합한 조합일 수 있다.
- [0044] 통신 버스(150)는 프로세서(110), 컴퓨터 판독 가능한 저장 매체(130)를 포함하여 게이트-단위 가지치기 장치(100)의 다른 다양한 컴포넌트들을 상호 연결한다.
- [0045] 게이트-단위 가지치기 장치(100)는 또한 하나 이상의 입출력 장치를 위한 인터페이스를 제공하는 하나 이상의 입출력 인터페이스(170) 및 하나 이상의 통신 인터페이스(190)를 포함할 수 있다. 입출력 인터페이스(170) 및 통신 인터페이스(190)는 통신 버스(150)에 연결된다. 입출력 장치(도시하지 않음)는 입출력 인터페이스(170)를 통해 게이트-단위 가지치기 장치(100)의 다른 컴포넌트들에 연결될 수 있다.
- [0048] 그러면, 도 2 및 도 3을 참조하여 본 발명의 바람직한 실시예에 따른 LSTM 가속을 위한 게이트-단위 가지치기 방법에 대하여 설명한다.
- [0049] 도 2는 본 발명의 바람직한 실시예에 따른 LSTM 가속을 위한 게이트-단위 가지치기 방법을 설명하기 흐름도이고, 도 3은 본 발명의 바람직한 실시예에 따른 LSTM의 게이트를 설명하기 위한 도면이다.
- [0050] 도 2를 참조하면, 게이트-단위 가지치기 장치(100)의 프로세서(110)는 시계열 데이터셋(dataset)을 이용하여 LSTM을 포함하는 모델을 훈련할 수 있다(S110).
- [0051] 즉, 프로세서(110)는 시계열 데이터셋을 이용한 모델의 훈련 과정을 통해, LSTM 내부의 가중치 값을 업데이트할 수 있다.



- [0052] 여기서, LSTM은 RNN(Recurrent Neural Network)의 장기 의존성(Long-Term Dependency) 문제, 즉 입력 시퀀스의 길이가 길면 시간이 흐를수록 앞선 정보가 현재 상태(state)에 제대로 전달되지 못하는 현상을 해결할 수 있다. 도 3을 참조하면, 각 단계에서의 정보를 셀 상태(cell state)에 저장하여 다음 단계에 보내면, 이전 단계의 내용을 얼마나 잊을지 정하고(망각 게이트), 현재의 입력 정보를 얼마나 받아들일지 정한 후(입력 게이트), 이 정보들을 이용하여 현재 단계에서 저장할 셀 상태를 결정하며(입력 업데이트 게이트), 히든 상태(hidden state)로 출력할 정보를 결정(출력 게이트)한다. 아래의 수식은 각 게이트에서의 연산을 식으로 나타낸 것이며, 주로 사용되는 연산인  $W_{xG}x_t$  또는  $W_{hG}h_{t-1}$ (G는 각 게이트)를 살펴보면 행렬-벡터 곱을 총 8번 연산하는 것을 확인할 수 있다.
- [0053] - (망각 게이트)  $f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f)$
- [0054] - (입력 게이트)  $i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i)$
- [0055] - (입력 업데이트 게이트)  $\tilde{c}_t = \tanh(W_{xu}x_t + W_{hu}h_{t-1} + b_u)$
- [0056] - (출력 게이트)  $o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o)$
- [0057] - (셀 상태)  $c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$
- [0058] - (히든 상태)  $h_t = o_t \odot \tanh(c_t)$
- [0060] 그런 다음, 프로세서(110)는 훈련된 모델을 기반으로, LSTM의 게이트별로 가지치기를 수행할 수 있다(S120).
- [0061] 즉, 프로세서(110)는 LSTM의 망각 게이트, LSTM의 입력 게이트, LSTM의 입력 업데이트 게이트 및 LSTM의 출력 게이트 각각에 대한 가지치기를 수행할 수 있다. 예컨대, 프로세서(110)는 LSTM 내부의 각 게이트별 가중치에서 결과에 영향이 적은 불필요한 가중치를 제거하여 게이트 단위의 가지치기를 수행할 수 있다. 일반적으로 LSTM은 망각 게이트, 입력 게이트 및 출력 게이트와 같은 총 3개의 게이트 구조를 가지고 있다. 그러나, 도 3에 도시된 바와 같이, LSTM의 입력 게이트는 2개의 게이트(입력 게이트 및 입력 업데이트 게이트)가 포함되어 있으므로, 본 발명은 망각 게이트, 입력 게이트, 입력 업데이트 게이트 및 출력 게이트와 같은 총 4개의 게이트로 분리하여 가지치기를 수행한다.
- [0062] 이때, 프로세서(110)는 미리 설정된 가지치기 비율을 기반으로, 망각 게이트, 입력 게이트, 입력 업데이트 게이트 및 출력 게이트 각각에 대한 가지치기를 수행할 수 있다.
- [0063] 여기서, 프로세서(110)는 망각 게이트, 입력 게이트, 입력 업데이트 게이트 및 출력 게이트 각각에 대한 가지치기를 서로 동일한 가지치기 비율을 기반으로 수행할 수 있다.
- [0065] 이후, 프로세서(110)는 가지치기가 수행된 모델을 미세-조정(fine-tuning)할 수 있다.
- [0068] 그러면, 도 4 내지 도 7을 참조하여 본 발명의 바람직한 실시예에 따른 가지치기 수행 단계에 대하여 보다 자세히 설명한다.
- [0069] 도 4는 도 2에 도시한 가지치기 수행 단계를 설명하기 위한 흐름도이고, 도 5는 본 발명의 바람직한 실시예에 따른 LSTM의 게이트별 가중치 행렬을 설명하기 위한 도면이며, 도 6은 본 발명의 바람직한 실시예에 따른 열 단위 가지치기를 설명하기 위한 도면이고, 도 7은 본 발명의 바람직한 실시예에 따른 LSTM의 내부 연산 순서를 설명하기 위한 도면이다.
- [0070] 도 4를 참조하면, 게이트-단위 가지치기 장치(100)의 프로세서(110)는 게이트 각각에 대한 하나의 가중치 행렬을 획득할 수 있다(S121).
- [0071] 즉, 프로세서(110)는 망각 게이트, 입력 게이트, 입력 업데이트 게이트 및 출력 게이트 각각에 대해, 게이트에 대한 복수개의 행렬-벡터 곱 연산을 통합(concatenate)하여 하나의 행렬-벡터 곱 연산에 대한 하나의 가중치 행렬을 획득할 수 있다.
- [0072] 보다 자세히 설명하면, 게이트 단위로 구조화 가지치기를 수행하기 위해, 게이트 단위로 연산을 통합해주는 과정을 진행한다. 각 게이트에는 히든 상태의 벡터와 입력의 벡터에 각각 가중치 행렬을 곱해준다. 각 게이트에는 2개의 벡터(히든 상태 벡터와 입력 벡터)가 입력으로 들어오기 때문에, 각 게이트는 총 2개의 행렬-벡터 곱 연산이 진행되어, 4개의 게이트는 총 8개의 행렬-벡터 곱 연산이 진행된다. 본 발명은 도 5에 도시된 바와 같

이, 게이트 단위로 가지치기를 수행하기 때문에, 각 게이트마다의 2개의 행렬-벡터 곱 연산을 통합하여 하나의 행렬-벡터 곱 연산을 만들게 된다. 이에 따라, LSTM의 셀 내부에는 총 4개의 행렬-벡터 곱 연산이 존재하게 된다.

- [0074] 그런 다음, 프로세서(110)는 게이트 각각에 대한 하나의 가중치 행렬을 기반으로 게이트 각각에 대한 가지치기를 수행할 수 있다(S122).
- [0075] 즉, 프로세서(110)는 망각 게이트, 입력 게이트, 입력 업데이트 게이트 및 출력 게이트 각각에 대해 획득한 하나의 가중치 행렬을 기반으로 가지치기를 수행할 수 있다.
- [0076] 보다 자세하게 설명하면, 각 게이트별로 구조화 가지치기를 수행한다. 구조화 가지치기는 가지치기 이후에도 가중치 행렬이 구조적 특징을 유지하도록 하기 위해 가중치 행렬의 열 벡터나 행 벡터 단위로 가지치기를 진행하는 것을 말한다. 도 5에 도시된 바와 같이, 벡터 단위로 제거를 하면 가중치 행렬은 기존의 기존의 텐스(dense)한 행렬 곱 연산을 그대로 사용할 수 있다. 각 게이트별로 구조화 가지치기를 수행하는 이유는 4개의 게이트를 통합하여 1개의 가중치 행렬로 만들어 해당 가중치 행렬에 구조화 가지치기를 수행하면 가지치기를 하기 위해 제거하는 최소 단위인 가지치기 유닛(pruning unit)의 크기가 수백에서 수천에 이르기 때문이다. 이때, 가지치기 유닛의 크기가 크면 하나의 가지치기 유닛을 제거할 때 발생하는 데이터 손실이 커지게 된다. 하지만, 4개의 게이트로 나누어 구조화 가지치기를 수행하면 가지치기 유닛의 크기가 4분의 1로 줄어들기 때문에 이와 비례하여 데이터 손실도 감소하는 효과를 얻을 수 있다.
- [0077] 이때, 프로세서(110)는 하나의 가중치 행렬을 기반으로, 열(column) 단위의 가지치기와 행(row) 단위의 가지치기를 수행할 수 있다.
- [0078] 즉, 프로세서(110)는 하나의 가중치 행렬을 기반으로, 각 열 벡터(column vector)의 절대값 평균을 획득하고, 가지치기 비율이 되도록 절대값 평균이 작은 순서대로 열 벡터를 제거하여 열 단위의 가지치기를 수행할 수 있다.
- [0079] 그리고, 프로세서(110)는 하나의 가중치 행렬을 기반으로, 각 행 벡터(row vector)의 절대값 평균을 획득하고, 가지치기 비율이 되도록 절대값 평균이 작은 순서대로 행 벡터를 제거하여 행 단위의 가지치기를 수행할 수 있다.
- [0080] 보다 자세하게 설명하면, 도 6에 도시된 바와 같이, 하나의 가중치 행렬(히든 상태에 대한 가중치 행렬과 입력에 대한 가중치 행렬을 통합한 행렬)을 열 벡터 단위로 나눈다. 그 후에 각 열 벡터의 절대값 평균을 계산하여 미리 설정된 가지치기 비율이 되도록 절대값 평균이 작은 순서대로 가지치기를 수행한다. 행 단위의 가지치기도 열 단위의 가지치기와 동일하게, 각 행 벡터의 절대값 평균을 계산하여 미리 설정된 가지치기 비율이 되도록 절대값 평균이 작은 순서대로 가지치기를 수행한다. 이와 같은, 열 단위의 가지치기와 행 단위의 가지치기를 각 게이트별로 수행하여, 게이트 단위의 가지치기를 수행할 수 있다.
- [0081] 여기서, 프로세서(110)는 망각 게이트, 입력 게이트, 입력 업데이트 게이트 및 출력 게이트 각각에 대한 가지치기를 서로 동일한 가지치기 비율을 기반으로 수행할 수 있다.
- [0082] 보다 자세하게 설명하면, 게이트 단위로 구조화 가지치기를 수행하면, 각 게이트별로 상이한 가지치기 비율을 갖게 되어 불균형(imbalance) 문제가 발생한다. 도 7에 도시된 바와 같은 LSTM의 내부 연산 순서를 보면, 앞선 연산의 작업부하(workload)가 다를 경우, 다음 순서의 모든 연산은 앞선 연산이 끝날 때까지 기다려야 한다. 이와 같은 문제를 해결하기 위해, 본 발명은 게이트 간에 서로 동일한 가지치기 비율을 설정하여 게이트 단위의 가지치기를 수행한다. 게이트별로 구조화 가지치기를 수행하는데 있어 가지치기 비율을 사용자에게 의해 선택받고, 이에 따라 게이트 내의 가중치 행렬은 해당하는 가지치기 비율이 될 때까지 벡터의 대표 값이 작은 순으로 제거한다. 이렇게 되면 모든 게이트는 서로 동일한 가지치기 비율을 갖는 가중치 행렬을 갖게 되고, Element-Wise 곱셈과 덧셈을 하는데 있어 모든 게이트가 동일한 속도로 처리되기 때문에 특정 게이트를 기다리지 않아도 된다. 이렇게 되면 LSTM을 처리하는 하드웨어의 효율(Utilization)을 극대화할 수 있다.
- [0084] 이와 같이, 본 발명은 LSTM의 가지치기를 게이트 단위로 수행하기 때문에 LSTM에 요구되는 메모리 사용량을 줄일 수 있다. 그리고, 본 발명은 게이트 수준으로 구조화 가지치기를 수행하기 때문에 종래의 구조화 가지치기 대비 높은 가지치기 비율을 갖을 수 있다. 또한, 본 발명은 게이트별로 동일한 가지치기 비율을 갖기 때문에 게이트 간의 지연 시간(delay time)이 발생하지 않아 하드웨어 효율을 극대화할 수 있다.
- [0087] 본 실시예들에 따른 동작은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨

터 판독 가능한 저장 매체에 기록될 수 있다. 컴퓨터 판독 가능한 저장 매체는 실행을 위해 프로세서에 명령어를 제공하는데 참여한 임의의 매체를 나타낸다. 컴퓨터 판독 가능한 저장 매체는 프로그램 명령, 데이터 파일, 데이터 구조 또는 이들의 조합을 포함할 수 있다. 예컨대, 자기 매체, 광기록 매체, 메모리 등이 있을 수 있다. 컴퓨터 프로그램은 네트워크로 연결된 컴퓨터 시스템 상에 분산되어 분산 방식으로 컴퓨터가 읽을 수 있는 코드가 저장되고 실행될 수도 있다. 본 실시예를 구현하기 위한 기능적인(Functional) 프로그램, 코드, 및 코드 세그먼트들은 본 실시예가 속하는 기술 분야의 프로그래머들에 의해 용이하게 추론될 수 있을 것이다.

[0088] 본 실시예들은 본 실시예의 기술 사상을 설명하기 위한 것이고, 이러한 실시예에 의하여 본 실시예의 기술 사상의 범위가 한정되는 것은 아니다. 본 실시예의 보호 범위는 아래의 청구범위에 의하여 해석되어야 하며, 그와 동등한 범위 내에 있는 모든 기술 사상은 본 실시예의 권리범위에 포함되는 것으로 해석되어야 할 것이다.

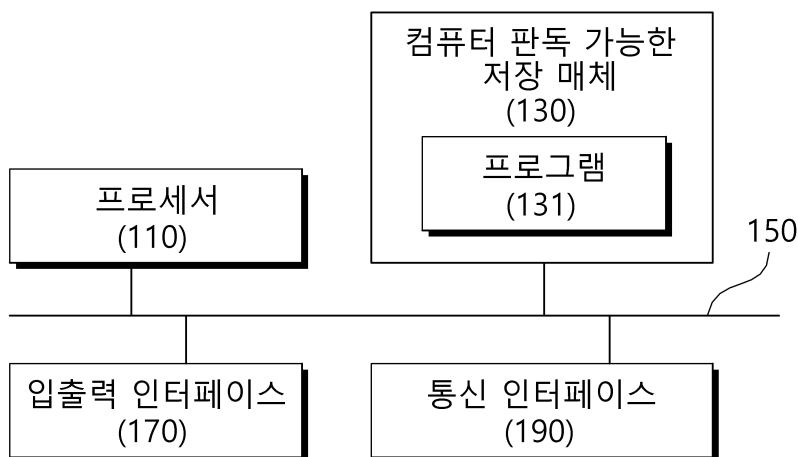
### 부호의 설명

[0089] 100 : 게이트-단위 가지치기 장치,  
110 : 프로세서,  
130 : 컴퓨터 판독 가능한 저장 매체,  
131 : 프로그램,  
150 : 통신 버스,  
170 : 입출력 인터페이스,  
190 : 통신 인터페이스

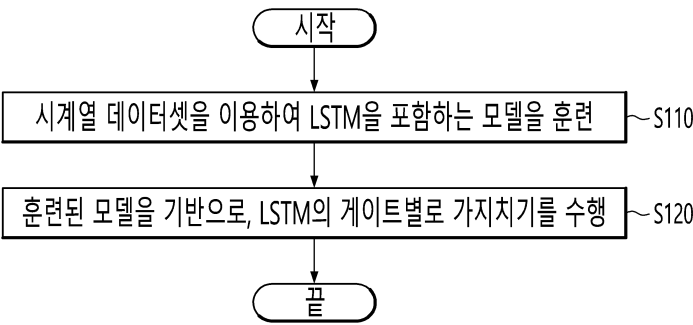
### 도면

#### 도면1

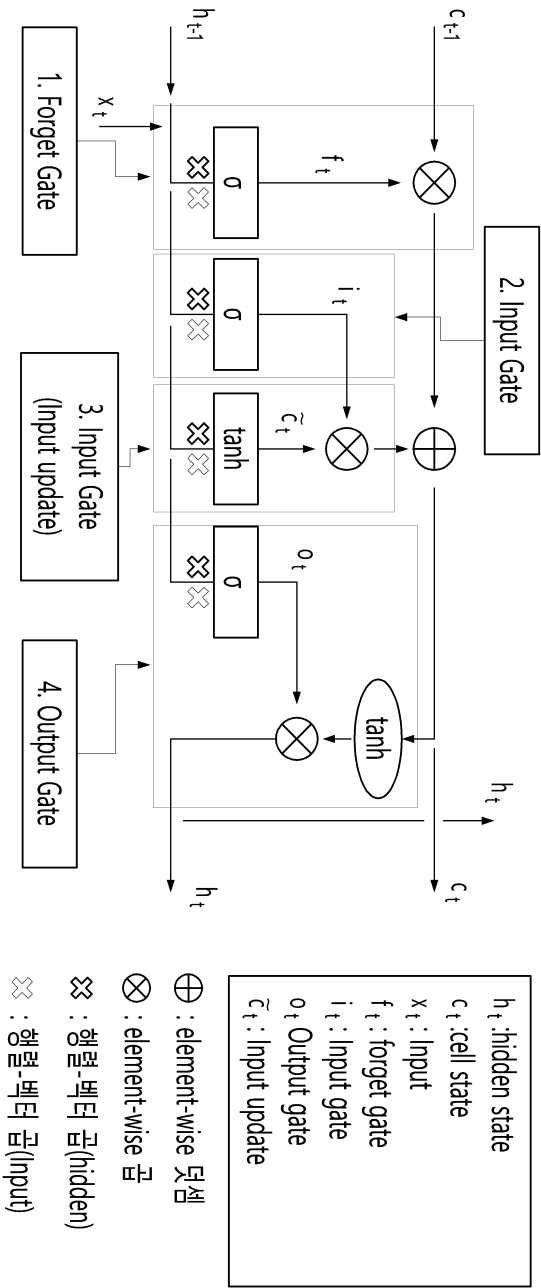
100



도면2

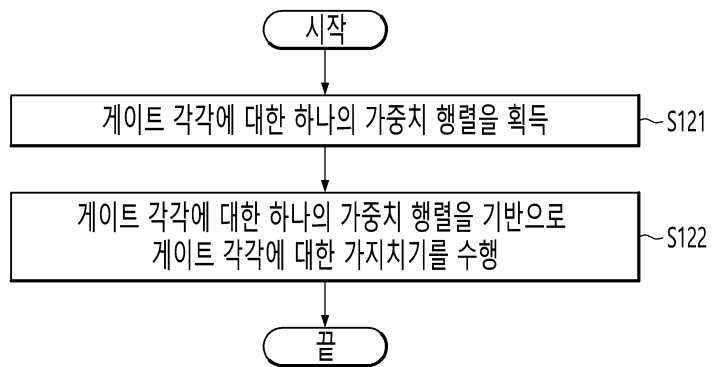


도면3

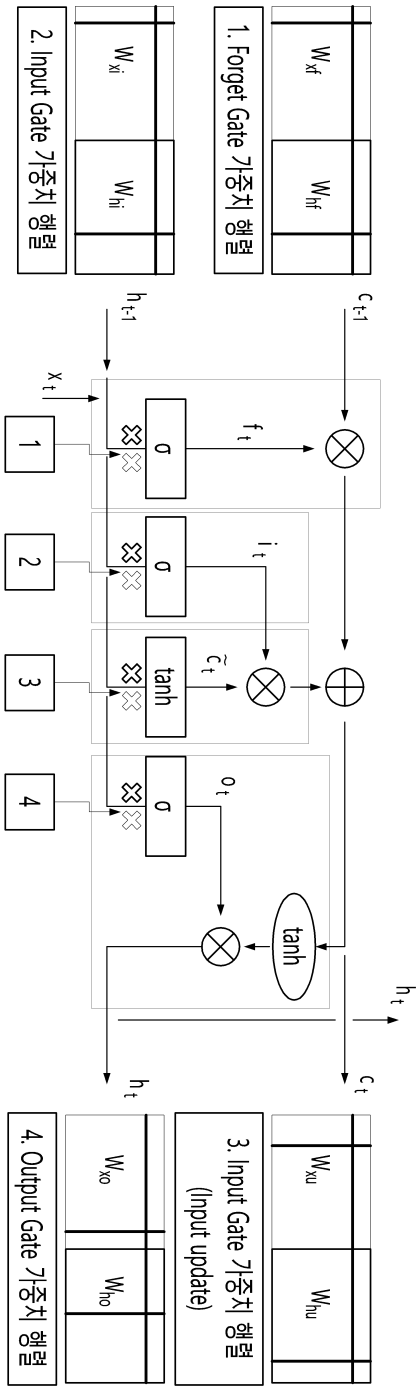


도면4

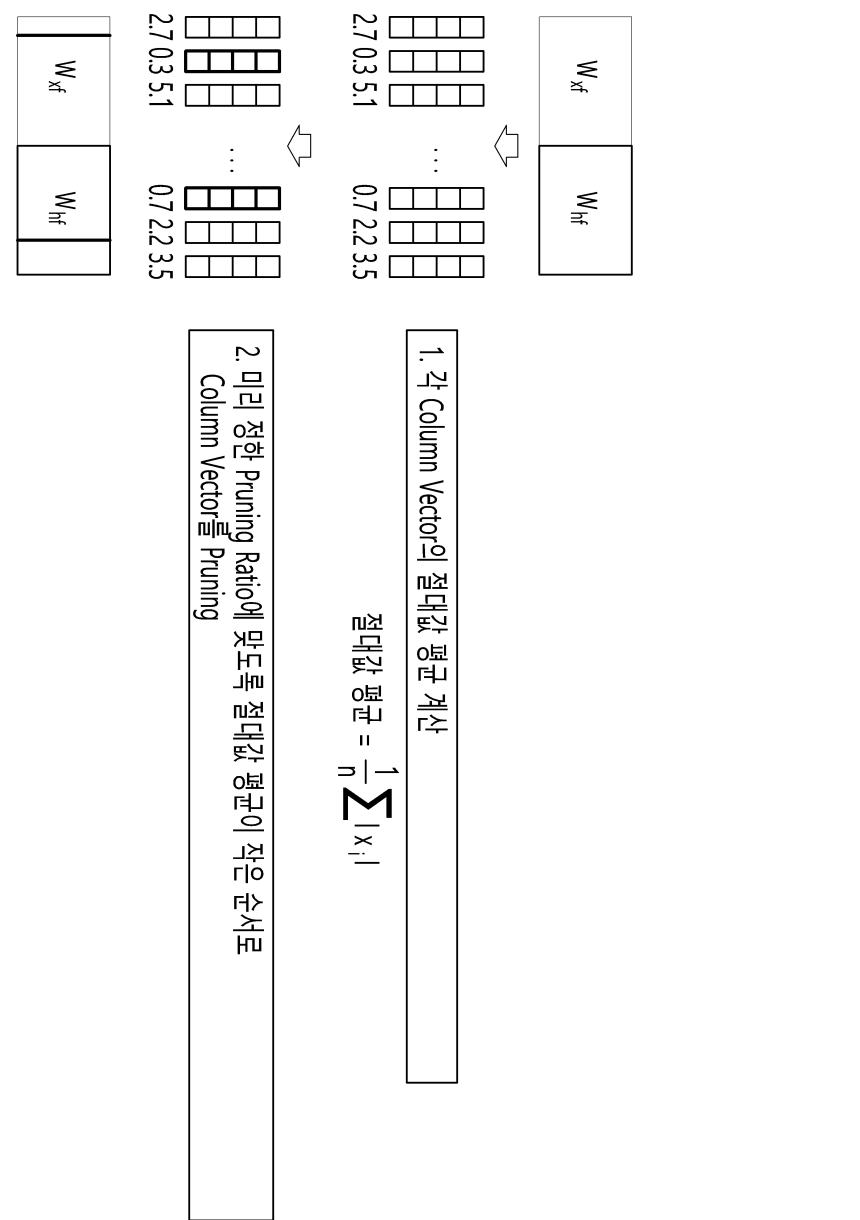
S120



도면5



도면6



도면7

