(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0032799
(43) 공개일자 2022년03월15일

(51) 국제특허분류(Int. Cl.)
G06N 3/04 (2006.01) *G06N 3/063* (2006.01)
G06N 3/08 (2006.01)
(52) CPC특허분류
G06N 3/04 (2013.01)
G06N 3/063 (2013.01)
(21) 출원번호 10-2020-0114564
(22) 출원일자 2020년09월08일
심사청구일자 없음

(71) 출원인
삼성전자주식회사
경기도 수원시 영통구 삼성로 129 (매탄동)
(72) 발명자
지형탁
경기도 화성시 동탄순환대로17길 15 ,2403-1305(중흥에스클래스에듀하이)
(74) 대리인
특허법인 무한

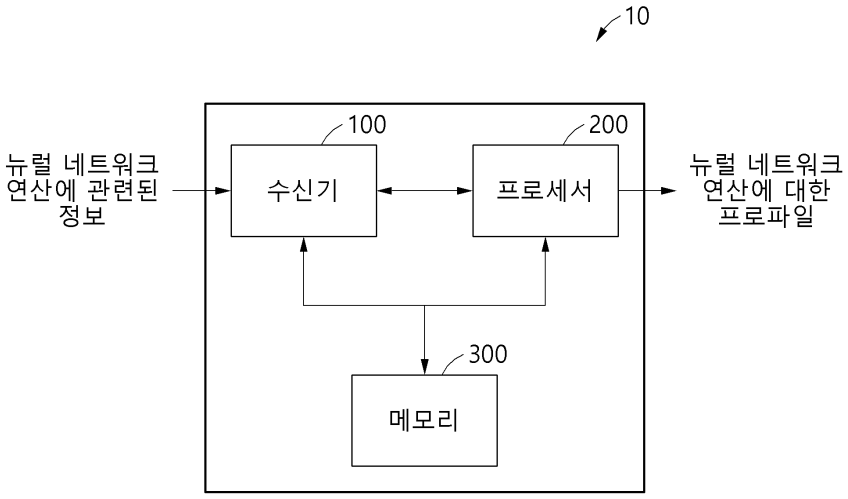
전체 청구항 수 : 총 19 항

(54) 발명의 명칭 **뉴럴 네트워크 프로파일링 방법 및 장치**

(57) 요 약

뉴럴 네트워크 프로파일링 방법 및 장치가 개시된다. 일 실시예에 따른 뉴럴 네트워크 프로파일링 방법은, 뉴럴 네트워크 연산에 관련된 이벤트(event) 및 상기 뉴럴 네트워크 연산을 수행하기 위한 제어 프로그램을 수신하는 단계와, 상기 이벤트 및 상기 제어 프로그램에 기초하여 미싱 이벤트(missing event)를 감지하는 단계와, 감지 결과에 기초하여 상기 뉴럴 네트워크 연산에 대한 프로파일(profile)을 생성하는 단계를 포함한다.

대 표 도 - 도1



(52) CPC특허분류
G06N 3/08 (2013.01)

명 세 서

청구범위

청구항 1

뉴럴 네트워크 연산에 관련된 이벤트(event) 및 상기 뉴럴 네트워크 연산을 수행하기 위한 제어 프로그램을 수신하는 단계;

상기 이벤트 및 상기 제어 프로그램에 기초하여 미싱 이벤트(missing event)를 감지하는 단계; 및

감지 결과에 기초하여 상기 뉴럴 네트워크 연산에 대한 프로파일(profile)을 생성하는 단계

를 포함하는 뉴럴 네트워크 프로파일링 방법.

청구항 2

제1항에 있어서,

상기 이벤트는,

상기 뉴럴 네트워크 연산의 시작 이벤트 및 종료 이벤트를 포함하는

뉴럴 네트워크 프로파일링 방법.

청구항 3

제1항에 있어서,

상기 제어 프로그램은,

상기 뉴럴 네트워크 연산의 실행 순서를 포함하는

뉴럴 네트워크 프로파일링 방법.

청구항 4

제1항에 있어서,

상기 감지하는 단계는,

상기 이벤트가 상기 제어 프로그램에 포함된 실행 순서에 매칭되는지 여부를 판단하는 단계; 및

판단 결과에 기초하여 상기 미싱 이벤트를 감지하는 단계

를 포함하는 뉴럴 네트워크 프로파일링 방법.

청구항 5

제1항에 있어서,

상기 생성하는 단계는,

상기 미싱 이벤트의 종류를 결정하는 단계; 및

상기 종류에 기초하여 상기 미싱 이벤트를 보완함으로써 상기 프로파일을 생성하는 단계

를 포함하는 뉴럴 네트워크 프로파일링 방법.

청구항 6

제5항에 있어서,

상기 종류에 기초하여 상기 미싱 이벤트를 보완함으로써 상기 프로파일을 생성하는 단계는,

상기 미싱 이벤트의 종류가 시작 이벤트인 경우, 상기 미싱 이벤트의 다음 이벤트로부터 제1 시간만큼 뺀 시간에 상기 시작 이벤트를 삽입하는 단계

를 포함하는 뉴럴 네트워크 프로파일링 방법.

청구항 7

제5항에 있어서,

상기 종류에 기초하여 상기 미싱 이벤트를 보완함으로써 상기 프로파일을 생성하는 단계는,

상기 미싱 이벤트의 종류가 종료 이벤트인 경우,

상기 뉴럴 네트워크 연산이 다른 연산과 관련된 이벤트와 오버랩되는지 여부를 판단하는 단계; 및

판단 결과에 기초하여 상기 종료 이벤트를 삽입하는 단계

를 포함하는 뉴럴 네트워크 프로파일링 방법.

청구항 8

제7항에 있어서,

상기 종료 이벤트를 삽입하는 단계는,

상기 연산이 상기 다른 연산과 관련된 이벤트와 오버랩되는 경우, 오버랩이 시작되는 부분에 상기 종료 이벤트를 삽입하는 단계

를 더 포함하는 뉴럴 네트워크 프로파일링 방법.

청구항 9

제7항에 있어서,

상기 종료 이벤트를 삽입하는 단계는,

상기 연산이 상기 다른 연산과 관련된 이벤트와 오버랩되지 않는 경우, 상기 미싱 이벤트의 다음 이벤트로부터 제2 시간만큼 뺀 시간에 상기 종료 이벤트를 삽입하는 단계

를 포함하는 뉴럴 네트워크 프로파일링 방법.

청구항 10

하드웨어와 결합되어 제1항 내지 제9 중 어느 하나의 항의 방법을 실행시키기 위하여 매체에 저장된 컴퓨터 프로그램.

청구항 11

뉴럴 네트워크 연산에 관련된 이벤트(event) 및 상기 뉴럴 네트워크 연산을 수행하기 위한 제어 프로그램을 수신하는 수신기; 및

상기 이벤트 및 상기 제어 프로그램에 기초하여 미싱 이벤트(missing event)를 감지하고, 감지 결과에 기초하여 상기 뉴럴 네트워크 연산에 대한 프로파일(profile)을 생성하는 프로세서

를 포함하는 뉴럴 네트워크 프로파일링 장치.

청구항 12

제11항에 있어서,

상기 이벤트는,

상기 뉴럴 네트워크 연산의 시작 이벤트 및 종료 이벤트를 포함하는

뉴럴 네트워크 프로파일링 장치.

청구항 13

제11항에 있어서,

상기 제어 프로그램은,

상기 뉴럴 네트워크 연산의 실행 순서를 포함하는

뉴럴 네트워크 프로파일링 장치.

청구항 14

제11항에 있어서,

상기 프로세서는,

상기 이벤트가 상기 제어 프로그램에 포함된 실행 순서에 매칭되는지 여부를 판단하고,

판단 결과에 기초하여 상기 미싱 이벤트를 감지하는

뉴럴 네트워크 프로파일링 장치.

청구항 15

제11항에 있어서,

상기 프로세서는,

상기 미싱 이벤트의 종류를 결정하고,

상기 종류에 기초하여 상기 미싱 이벤트를 보완함으로써 상기 프로파일을 생성하는

뉴럴 네트워크 프로파일링 장치.

청구항 16

제15항에 있어서,

상기 프로세서는,

상기 미싱 이벤트의 종류가 시작 이벤트인 경우, 상기 미싱 이벤트의 다음 이벤트로부터 제1 시간만큼 뺀 시간에 상기 시작 이벤트를 삽입하는

뉴럴 네트워크 프로파일링 장치.

청구항 17

제15항에 있어서,

상기 프로세서는,

상기 미싱 이벤트의 종류가 종료 이벤트인 경우,

상기 뉴럴 네트워크 연산이 다른 연산과 관련된 이벤트와 오버랩되는지 여부를 판단하고,

판단 결과에 기초하여 상기 종료 이벤트를 삽입하는

뉴럴 네트워크 프로파일링 장치.

청구항 18

제17항에 있어서,

상기 프로세서는,

상기 연산이 상기 다른 연산과 관련된 이벤트와 오버랩되는 경우, 오버랩이 시작되는 부분에 상기 종료 이벤트를 삽입하는

뉴럴 네트워크 프로파일링 장치.

청구항 19

제17항에 있어서,

상기 프로세서는,

상기 연산이 상기 다른 연산과 관련된 이벤트와 오버랩되지 않는 경우, 상기 미싱 이벤트의 다음 이벤트로부터 제2 시간만큼 뺀 시간에 상기 종료 이벤트를 삽입하는

뉴럴 네트워크 프로파일링 장치.

발명의 설명

기 술 분 야

[0001] 아래 실시예들은 뉴럴 네트워크 프로파일링 방법 및 장치에 관한 것이다.

배 경 기 술

[0002] 종래의 에뮬레이터 추론(emulator inference)의 경우, NPU(Neural Processing Unit)의 프로파일링을 수행하기 위해, NPU의 RTL(Register Transfer Level)을 에뮬레이터와 에뮬레티어를 실행하는 보드(board)에 업로드하여 추론을 수행하고, 추론이 끝난 후, 로그를 다운로드 하여 프로파일용 데이터로 파싱을 통해 프로파일링 수행하였다.

[0003] 또는, 타겟 추론(target inference)의 경우, 모바일 폰 커널 드라이버 단에서 NPU의 하드웨어 이벤트 신호(event signal)을 ARM STM(System Trace Macrocell)과 연결하여 추론시 이벤트 정보를 얻는다.

[0004] 종래의 방식은 데이터 후처리 과정이 필요하고, 로그 파일의 용량이 커 프로파일링을 수행하는데 긴 시간이 소요된다는 문제가 있다. 또한, 현재 추론 중인 뉴럴 네트워크에서 어느 부분이 수행되는지 파악하는 것이 어렵고, 이벤트 로그가 사라지면(missing) 프로파일링 데이터가 부정확해진다는 문제점을 갖는다.

발명의 내용

해결하려는 과제

과제의 해결 수단

[0005] 일 실시예에 따른 뉴럴 네트워크 프로파일링 방법은, 뉴럴 네트워크 연산에 관련된 이벤트(event) 및 상기 뉴럴 네트워크 연산을 수행하기 위한 제어 프로그램을 수신하는 단계와, 상기 이벤트 및 상기 제어 프로그램에 기초하여 미싱 이벤트(missing event)를 감지하는 단계와, 감지 결과에 기초하여 상기 뉴럴 네트워크 연산에 대한 프로파일(profile)을 생성하는 단계를 포함한다.

[0006] 상기 이벤트는, 상기 뉴럴 네트워크 연산의 시작 이벤트 및 종료 이벤트를 포함할 수 있다.

[0007] 상기 제어 프로그램은, 상기 뉴럴 네트워크 연산의 실행 순서를 포함할 수 있다.

[0008] 상기 감지하는 단계는, 상기 이벤트가 상기 제어 프로그램에 포함된 실행 순서에 매칭되는지 여부를 판단하는 단계와, 판단 결과에 기초하여 상기 미싱 이벤트를 감지하는 단계를 포함할 수 있다.

[0009] 상기 생성하는 단계는, 상기 미싱 이벤트의 종류를 결정하는 단계와, 상기 종류에 기초하여 상기 미싱 이벤트를 보완함으로써 상기 프로파일을 생성하는 단계를 포함할 수 있다.

[0010] 상기 종류에 기초하여 상기 미싱 이벤트를 보완함으로써 상기 프로파일을 생성하는 단계는, 상기 미싱 이벤트의 종류가 시작 이벤트인 경우, 상기 미싱 이벤트의 다음 이벤트로부터 제1 시간만큼 뺀 시간에 상기 시작 이벤트를 삽입하는 단계를 포함할 수 있다.

[0011] 상기 종류에 기초하여 상기 미싱 이벤트를 보완함으로써 상기 프로파일을 생성하는 단계는, 상기 미싱 이벤트의 종류가 종료 이벤트인 경우, 상기 뉴럴 네트워크 연산이 다른 연산과 관련된 이벤트와 오버랩되는지 여부를 판단하는 단계와, 판단 결과에 기초하여 상기 종료 이벤트를 삽입하는 단계를 포함할 수 있다.

[0012] 상기 종료 이벤트를 삽입하는 단계는, 상기 연산이 상기 다른 연산과 관련된 이벤트와 오버랩되는 경우, 오버랩이 시작되는 부분에 상기 종료 이벤트를 삽입하는 단계를 더 포함할 수 있다.

[0013] 상기 종료 이벤트를 삽입하는 단계는, 상기 연산이 상기 다른 연산과 관련된 이벤트와 오버랩되지 않는 경우, 상기 미싱 이벤트의 다음 이벤트로부터 제2 시간만큼 뺀 시간에 상기 종료 이벤트를 삽입하는 단계를 포함할 수 있다.

[0014] 일 실시예에 따른 뉴럴 네트워크 프로파일링 장치는, 뉴럴 네트워크 연산에 관련된 이벤트(event) 및 상기 뉴럴 네트워크 연산을 수행하기 위한 제어 프로그램을 수신하는 수신기와, 상기 이벤트 및 상기 제어 프로그램에 기초하여 미싱 이벤트(missing event)를 감지하고, 감지 결과에 기초하여 상기 뉴럴 네트워크 연산에 대한 프로파일(profile)을 생성하는 프로세서를 포함한다.

[0015] 상기 이벤트는, 상기 뉴럴 네트워크 연산의 시작 이벤트 및 종료 이벤트를 포함할 수 있다.

[0016] 상기 제어 프로그램은, 상기 뉴럴 네트워크 연산의 실행 순서를 포함할 수 있다.

[0017] 상기 프로세서는, 상기 이벤트가 상기 제어 프로그램에 포함된 실행 순서에 매칭되는지 여부를 판단하고, 판단 결과에 기초하여 상기 미싱 이벤트를 감지할 수 있다.

[0018] 상기 프로세서는, 상기 미싱 이벤트의 종류를 결정하고, 상기 종류에 기초하여 상기 미싱 이벤트를 보완함으로써 상기 프로파일을 생성할 수 있다.

[0019] 상기 프로세서는, 상기 미싱 이벤트의 종류가 시작 이벤트인 경우, 상기 미싱 이벤트의 다음 이벤트로부터 제1 시간만큼 뺀 시간에 상기 시작 이벤트를 삽입할 수 있다.

[0020] 상기 프로세서는, 상기 미싱 이벤트의 종류가 종료 이벤트인 경우, 상기 뉴럴 네트워크 연산이 다른 연산과 관

련된 이벤트와 오버랩되는지 여부를 판단하고, 판단 결과에 기초하여 상기 종료 이벤트를 삽입할 수 있다.

[0021] 상기 프로세서는, 상기 연산이 상기 다른 연산과 관련된 이벤트와 오버랩되는 경우, 오버랩이 시작되는 부분에 상기 종료 이벤트를 삽입할 수 있다.

[0022] 상기 프로세서는, 상기 연산이 상기 다른 연산과 관련된 이벤트와 오버랩되지 않는 경우, 상기 미싱 이벤트의 다음 이벤트로부터 제2 시간만큼 뺀 시간에 상기 종료 이벤트를 삽입할 수 있다.

도면의 간단한 설명

[0023] 도 1은 일 실시예에 따른 프로파일링 장치의 개략적인 블록도를 나타낸다.

도 2는 뉴럴 네트워크 처리 시스템의 개략적인 블록도를 나타낸다.

도 3은 도 1에 도시된 프로파일링 장치의 동작을 나타낸다.

도 4는 도 1에 도시된 프로파일링 장치가 미싱 이벤트를 보완하는 동작의 예를 나타낸다.

도 5는 도 1에 도시된 프로파일링 장치의 시각화 동작의 순서를 나타낸다.

도 6은 도 1에 도시된 프로파일링 장치의 동작의 순서를 나타낸다.

발명을 실시하기 위한 구체적인 내용

[0024] 이하에서, 첨부된 도면을 참조하여 실시예들을 상세하게 설명한다. 그러나, 실시예들에는 다양한 변경이 가해질 수 있어서 특허출원의 권리 범위가 이러한 실시예들에 의해 제한되거나 한정되는 것은 아니다. 실시예들에 대한 모든 변경, 균등물 내지 대체물이 권리 범위에 포함되는 것으로 이해되어야 한다.

[0025] 실시예에서 사용한 용어는 단지 설명을 목적으로 사용된 것으로, 한정하려는 의도로 해석되어서는 안된다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 명세서 상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0026] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 실시예가 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가지고 있다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥 상 가지는 의미와 일치하는 의미를 가지는 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0027] 또한, 첨부 도면을 참조하여 설명함에 있어, 도면 부호에 관계없이 동일한 구성 요소는 동일한 참조부호를 부여하고 이에 대한 중복되는 설명은 생략하기로 한다. 실시예를 설명함에 있어서 관련된 공지 기술에 대한 구체적인 설명이 실시예의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우 그 상세한 설명을 생략한다.

[0028] 또한, 실시 예의 구성 요소를 설명하는 데 있어서, 제 1, 제 2, A, B, (a), (b) 등의 용어를 사용할 수 있다. 이러한 용어는 그 구성 요소를 다른 구성 요소와 구별하기 위한 것일 뿐, 그 용어에 의해 해당 구성 요소의 본질이나 차례 또는 순서 등이 한정되지 않는다. 어떤 구성 요소가 다른 구성요소에 "연결", "결합" 또는 "접속"된다고 기재된 경우, 그 구성 요소는 그 다른 구성요소에 직접적으로 연결되거나 접속될 수 있지만, 각 구성 요소 사이에 또 다른 구성 요소가 "연결", "결합" 또는 "접속"될 수도 있다고 이해되어야 할 것이다.

[0029] 어느 하나의 실시 예에 포함된 구성요소와, 공통적인 기능을 포함하는 구성요소는, 다른 실시 예에서 동일한 명칭을 사용하여 설명하기로 한다. 반대되는 기재가 없는 이상, 어느 하나의 실시 예에 기재한 설명은 다른 실시 예에도 적용될 수 있으며, 중복되는 범위에서 구체적인 설명은 생략하기로 한다.

[0031] 도 1은 일 실시예에 따른 프로파일링(profiling) 장치의 개략적인 블록도를 나타낸다.

[0032] 도 1을 참조하면, 프로파일링 장치(10)는 뉴럴 네트워크 프로파일링을 수행할 수 있다. 프로파일링 장치(10)는 뉴럴 네트워크에서 수행되는 연산에 관한 프로파일링을 수행할 수 있다.

[0033] 프로파일링은 프로그램의 시간 복잡도 및 공간(메모리), 특정 명령어 이용, 함수 호출의 주기와 빈도 등을 측정

하는 동적 프로그램 분석의 한 형태를 의미할 수 있다. 프로파일링 정보는 뉴럴 네트워크의 최적화를 보조하기 위해 사용될 수 있다. 프로파일링 장치(10)는 프로그램 소스 코드나 이진 실행 파일을 계측 분석함으로써 프로파일링을 수행할 수 있다.

[0034] 프로파일(profile)은 프로파일링을 통해 생성된 데이터를 의미할 수 있다. 프로파일은 시간에 따라 수행되는 뉴럴 네트워크 연산에 관련된 이벤트를 나타낼 수 있다.

[0035] 뉴럴 네트워크(또는 인공 신경망)는 기계학습과 인지과학에서 생물학의 신경을 모방한 통계학적 학습 알고리즘을 포함할 수 있다. 뉴럴 네트워크는 시냅스의 결합으로 네트워크를 형성한 인공 뉴런(노드)이 학습을 통해 시냅스의 결합 세기를 변화시켜, 문제 해결 능력을 가지는 모델 전반을 의미할 수 있다.

[0036] 뉴럴 네트워크는 심층 뉴럴 네트워크 (Deep Neural Network)를 포함할 수 있다. 뉴럴 네트워크는 CNN(Convolutional Neural Network), RNN(Recurrent Neural Network), 퍼셉트론(perceptron), FF(Feed Forward), RBF(Radial Basis Network), DFF(Deep Feed Forward), LSTM(Long Short Term Memory), GRU(Gated Recurrent Unit), AE(Auto Encoder), VAE(Variational Auto Encoder), DAE(Denoising Auto Encoder), SAE(Sparse Auto Encoder), MC(Markov Chain), HN(Hopfield Network), BM(Boltzmann Machine), RBM(Restricted Boltzmann Machine), DBN(Depp Belief Network), DCN(Deep Convolutional Network), DN(Deconvolutional Network), DCIGN(Deep Convolutional Inverse Graphics Network), GAN(Generative Adversarial Network), LSM(Liquid State Machine), ELM(Extreme Learning Machine), ESN(Echo State Network), DRN(Deep Residual Network), DNC(Differentiable Neural Computer), NTM(Neural Turning Machine), CN(Capsule Network), KN(Kohonen Network) 및 AN(Attention Network)를 포함할 수 있다.

[0037] 프로파일링 장치(10)는 뉴럴 네트워크 연산과 관련된 이벤트에 기초하여 프로파일을 생성하고, 생성한 프로파일을 시각화(visualization)할 수 있다.

[0038] 프로파일링 장치(10)는 뉴럴 네트워크에 대한 프로파일을 생성함으로써 뉴럴 네트워크 모델을 추론하는 과정에서 예상한 하드웨어 사양에 적합한 계산 시간이 소요되는지, 예상한 사이클(cycle)에 맞게 뉴럴 네트워크 연산이 수행되는지를 검증할 수 있다. 또한, 프로파일링 장치(10)는 생성한 프로파일을 이용하여 뉴럴 네트워크의 최적화 지점(point)을 검출할 수 있다.

[0039] 프로파일링 장치(10)는 뉴럴 네트워크 연산에 관련된 정보를 처리하여 뉴럴 네트워크 연산에 대한 프로파일을 생성할 수 있다. 뉴럴 네트워크 연산에 관련된 정보는 뉴럴 네트워크 연산에 관련된 이벤트 및 뉴럴 네트워크 연산을 수행하기 위한 제어 프로그램을 포함할 수 있다.

[0040] 이벤트는 뉴럴 네트워크 연산의 종류에 따른 시작 과 종료를 나타낼 수 있다. 이벤트는 뉴럴 네트워크 연산의 시작 이벤트 및 종료 이벤트를 포함할 수 있다.

[0041] 제어 프로그램은 뉴럴 네트워크를 이용한 추론을 수행하기 위해 컴파일러에 의해 생성된 프로그램을 포함할 수 있다. 제어 프로그램은 뉴럴 네트워크 연산자 인트린직(intrinsic) 순서를 포함할 수 있다. 인트린직은 뉴럴 네트워크 연산을 수행하는 NPU 내부 함수(built in function)를 포함할 수 있다. 예를 들어, 제어 프로그램은 뉴럴 네트워크 연산의 실행 순서를 포함할 수 있다.

[0042] 프로파일링 장치(10)는 수신기(100) 및 프로세서(200)를 포함한다. 프로파일링 장치(10)는 메모리(300)를 더 포함할 수 있다.

[0043] 수신기(100)는 뉴럴 네트워크 연산에 관련된 이벤트(event) 및 뉴럴 네트워크 연산을 수행하기 위한 제어 프로그램(control program)을 수신할 수 있다.

[0044] 수신기(100)는 수신한 이벤트 및 제어 프로그램을 프로세서(200)로 출력할 수 있다. 수신기(100)는 수신 인터페이스를 포함할 수 있다.

[0045] 프로세서(200)는 메모리(300)에 저장된 데이터를 처리할 수 있다. 프로세서(200)는 메모리(300)에 저장된 컴퓨터로 읽을 수 있는 코드(예를 들어, 소프트웨어) 및 프로세서(200)에 의해 유발된 인스트럭션(instruction)들을 실행할 수 있다.

[0046] "프로세서(200)"는 목적하는 동작들(desired operations)을 실행시키기 위한 물리적인 구조를 갖는 회로를 가지는 하드웨어로 구현된 데이터 처리 장치일 수 있다. 예를 들어, 목적하는 동작들은 프로그램에 포함된 코드(code) 또는 인스트럭션들(instructions)을 포함할 수 있다.

[0047] 예를 들어, 하드웨어로 구현된 데이터 처리 장치는 마이크로프로세서(microprocessor), 중앙 처리 장치(central processing unit), 프로세서 코어(processor core), 멀티-코어 프로세서(multi-core processor), 멀티프로세서(multiprocessor), ASIC(Application-Specific Integrated Circuit), FPGA(Field Programmable Gate Array)를 포함할 수 있다.

[0048] 프로세서(200)는 이벤트 및 제어 프로그램에 기초하여 미싱 이벤트(missing event)를 감지할 수 있다. 미싱 이벤트는 제어 프로그램의 인트린직 상에는 포함되어 뉴럴 네트워크의 처리 과정에 있어 수행되었어야 하는데 수신한 이벤트에 포함되지 않은 이벤트를 의미할 수 있다.

[0049] 프로세서(200)는 이벤트가 제어 프로그램에 포함된 실행 순서에 매칭되는지 여부를 판단할 수 있다. 프로세서(200)는 판단 결과에 기초하여 미싱 이벤트를 감지할 수 있다.

[0050] 프로세서(200)는 미싱 이벤트의 감지 결과에 기초하여 뉴럴 네트워크 연산에 대한 프로파일(profile)을 생성할 수 있다. 프로세서(200)는 미싱 이벤트의 종류를 결정할 수 있다. 프로세서(200)는 결정된 종류에 기초하여 미싱 이벤트를 보완함으로써 프로파일을 생성할 수 있다.

[0051] 프로세서(200)는 미싱 이벤트의 종류가 시작 이벤트인 경우, 미싱 이벤트의 다음 이벤트로부터 제1 시간만큼 뺀 시간에 시작 이벤트를 삽입(insert)할 수 있다.

[0052] 프로세서(200)는 미싱 이벤트의 종류가 종료 이벤트인 경우, 뉴럴 네트워크 연산이 다른 연산과 관련된 이벤트와 오버랩되는지 여부를 판단할 수 있다. 프로세서(200)는 판단 결과에 기초하여 종료 이벤트를 삽입할 수 있다.

[0053] 프로세서(200)는 연산이 다른 연산과 관련된 이벤트와 오버랩되는 경우, 오버랩이 시작되는 부분에 종료 이벤트를 삽입할 수 있다. 프로세서(200)는 연산이 다른 연산과 관련된 이벤트와 오버랩되지 않는 경우, 미싱 이벤트의 다음 이벤트로부터 제2 시간만큼 뺀 시간에 종료 이벤트를 삽입할 수 있다.

[0054] 메모리(300)는 프로세서(200)에 의해 실행가능한 인스트럭션들(또는 프로그램)을 저장할 수 있다. 예를 들어, 인스트럭션들은 프로세서(200)의 동작 및/또는 프로세서(200)의 각 구성의 동작을 실행하기 위한 인스트럭션들을 포함할 수 있다.

[0055] 메모리(300)는 휘발성 메모리 장치 또는 불휘발성 메모리 장치로 구현될 수 있다.

[0056] 휘발성 메모리 장치는 DRAM(dynamic random access memory), SRAM(static random access memory), T-RAM(thyristor RAM), Z-RAM(zero capacitor RAM), 또는 TTRAM(Twin Transistor RAM)으로 구현될 수 있다.

[0057] 불휘발성 메모리 장치는 EEPROM(Electrically Erasable Programmable Read-Only Memory), 플래시(flash) 메모리, MRAM(Magnetic RAM), 스핀전달토크 MRAM(Spin-Transfer Torque(STT)-MRAM), Conductive Bridging RAM(CBRAM), FeRAM(Ferroelectric RAM), PRAM(Phase change RAM), 저항 메모리(Resistive RAM(RRAM)), 나노튜브 RRAM(Nanotube RRAM), 폴리머 RAM(Polymer RAM(PoRAM)), 나노 부유 게이트 메모리(Nano Floating Gate Memory(NFGM)), 홀로그래픽 메모리(holographic memory), 분자 전자 메모리 소자(Molecular Eelectronic Memory Device), 또는 절연 저항 변화 메모리(Insulator Resistance Change Memory)로 구현될 수 있다.

[0059] 도 2는 뉴럴 네트워크 처리 시스템의 개략적인 블록도를 나타낸다.

[0060] 도 2를 참조하면, 뉴럴 네트워크 처리 시스템은 프로파일링 장치(10)와 시스템 IP(Intellectual Property)는 서로 뉴럴 네트워크 연산에 관련된 정보를 송수신할 수 있다. 시스템 IP는 디버깅(debug) 및 성능(performance) 측정을 수행할 수 있다. 예를 들어, 시스팀 IP는 코어사이트(CoreSight)를 포함할 수 있다.

[0061] 프로파일링 장치(10)는 프로세서(200) 및 메모리(300)를 포함할 수 있고, 연산기(400)를 더 포함할 수 있다. 메모리(300)는 DRAM으로 구현될 수 있다. 메모리(300)는 트레이스 데이터(trace data)를 저장할 수 있다.

[0062] 연산기는 프로파일링 장치(10)의 내부 또는 외부에 구현될 수 있다.

[0063] 연산기(400)는 NPU(Neural Processing Unit) 또는 DSP(Digital Signal Processor)를 포함할 수 있다. 연산기(400)는 병합기를 포함할 수 있다. 병합기는 이벤트를 미리 정의할 수 있다. 병합기는 미리 정의된 세트 중 하나를 따라 이벤트를 결합할 수 있다.

[0064] 프로세서(200)는 연산기(400)로부터 뉴럴 네트워크 연산에 관련된 이벤트를 수신할 수 있다. 프로세서(200)는 수신한 이벤트와 제어 프로그램의 비교를 통해 미싱 이벤트를 보완함으로써 뉴럴 네트워크 프로파일을 생성할 수 있다.

[0066] 도 3은 도 1에 도시된 프로파일링 장치의 동작을 나타낸다.

[0067] 도 3을 참조하면, 프로파일링 장치(10)는 호스트(host) 장치에 구현될 수 있다. 예를 들어, 호스트 장치는 PC(Personal Computer) 또는 서버(server)로 구현될 수 있다. 프로파일링 장치(10)는 타겟 장치에서 수행되는 연산에 관련된 이벤트 정보를 수신하여 뉴럴 네트워크 연산에 대한 프로파일링을 수행할 수 있다.

[0068] 호스트 장치는 컴파일러를 포함할 수 있다. 컴파일러는 뉴럴 네트워크 빌딩을 수행할 수 있다(310). 컴파일러는 제어 프로그램을 생성할 수 있다(320). 예를 들어, 컴파일러는 NPU 용 실행 파일인 NCP(Network Control Program)를 생성할 수 있다.

[0069] 타겟 장치는 뉴럴 네트워크를 이용한 추론이 수행되는 장치를 의미할 수 있다. 예를 들어, 타겟 장치는 IoT 장치, Machine-type 통신 장치 또는 휴대용 전자 장치 등으로 구현될 수 있다.

[0070] 휴대용 전자 장치는 랩탑(laptop) 컴퓨터, 이동 전화기, 스마트 폰(smart phone), 태블릿(tablet) PC, 모바일 인터넷 디바이스(mobile internet device(MID)), PDA(personal digital assistant), EDA(enterprise digital assistant), 디지털 스틸 카메라(digital still camera), 디지털 비디오 카메라(digital video camera), PMP(portable multimedia player), PND(personal navigation device 또는 portable navigation device), 휴대용 게임 콘솔(handheld game console), e-북(e-book), 스마트 디바이스(smart device)로 구현될 수 있다. 예를 들어, 스마트 디바이스는 스마트 워치(smart watch) 또는 스마트 밴드(smart band)로 구현될 수 있다.

[0071] 타겟 장치는 NPU를 포함할 수 있다. 타겟 장치는 뉴럴 네트워크에 포함된 연산을 수행하는 NPU를 이용하여 추론을 수행할 수 있다(330). 타겟 장치는 추론을 수행하면서 이벤트 정보를 생성할 수 있다(340).

[0072] 수신기(100)는 제어 프로그램과 이벤트 정보를 수신할 수 있다. 프로세서(200)는 제어 프로그램과 이벤트 정보에 기초하여 뉴럴 네트워크 프로파일링을 수행할 수 있다(350). 프로파일링을 수행하는 과정은 도 4를 참조하여 보다 자세하게 설명한다.

[0073] 프로세서(200)는 생성한 프로파일에 기초하여 시각화를 수행할 수 있다(360).

[0075] 도 4는 도 1에 도시된 프로파일링 장치가 미싱 이벤트를 보완하는 동작의 예를 나타낸다.

[0076] 도 4를 참조하면, 프로세서(200)는 이벤트 및 제어 프로그램에 기초하여 미싱 이벤트(missing event)를 감지할 수 있다. 컴파일러는 제어 프로그램을 생성하여 전송할 수 있다. 예를 들어, 제어 프로그램은 NCP를 포함할 수 있다.

[0077] NCP는 그룹(group)을 실행 단위로 가질 수 있다. 그룹을 통해 네트워크의 수행 지점이 추정될 수 있다.

[0078] 도 4의 예시에서, 컴파일러가 생성한 NCP(인트린직)은 생성한 뉴럴 네트워크 연산의 실행 순서를 포함할 수 있다. 이벤트 정보는 NPU가 생성하여 전송한 이벤트 정보를 의미할 수 있다. 이벤트 정보는 파일의 형태로 수신될 수 있다.

[0079] 이벤트 정보와 제어 프로그램은 뉴럴 네트워크 연산과, 연산에 관련된 이벤트를 포함할 수 있다. 예를 들어, File은 컨볼루션(convolution) 연산을 의미하고, PU는 패드/풀(pad/pool) 연산을 의미하고, RU는 리포맷(reformat) 연산을 의미할 수 있다. 각 연산은 시작 이벤트와 종료 이벤트를 가질 수 있다.

[0080] 프로세서(200)는 이벤트가 제어 프로그램에 포함된 실행 순서에 매칭되는지 여부를 판단할 수 있다. 프로세서(200)는 판단 결과에 기초하여 미싱 이벤트를 감지할 수 있다.

[0081] 이 때, 프로세서(200)는 File, PU 및 RU 연산 각각은 서로 동시에 수행되지 않는 것을 전제로 미싱 이벤트를 감지할 수 있다.

[0082] 프로세서(200)는 미싱 이벤트의 감지 결과에 기초하여 뉴럴 네트워크 연산에 대한 프로파일(profile)을 생성할 수 있다. 프로세서(200)는 미싱 이벤트의 종류를 결정할 수 있다. 프로세서(200)는 결정된 종류에 기초하여

미싱 이벤트를 보완함으로써 프로파일을 생성할 수 있다.

[0083] 프로세서(200)는 미싱 이벤트의 종류가 시작 이벤트인 경우, 미싱 이벤트의 다음 이벤트로부터 제1 시간만큼 뺀 시간에 시작 이벤트를 삽입(insert)할 수 있다.

[0084] 미싱 이벤트의 종류가 시작 인벤트인 경우, 시작 이벤트 전에 DMA(Direct Memory Access)를 하면서 시작 이벤트가 발생하는지, 또는 다른 연산기의 이벤트가 끝나자 마자 시작 이벤트가 진행되는지 알 수 없기 때문에, 프로세서(200)는 종료 이벤트로부터 제1 시간만큼 뺀 시간에 시작 이벤트를 삽입할 수 있다.

[0085] 제1 시간은 연산의 종류와 하드웨어에 따라 상이할 수 있다. 예를 들어, 제1 시간은 10 ns일 수 있다.

[0086] 프로세서(200)는 미싱 이벤트의 종류가 종료 이벤트인 경우, 뉴럴 네트워크 연산이 다른 연산과 관련된 이벤트와 오버랩되는지 여부를 판단할 수 있다. 프로세서(200)는 판단 결과에 기초하여 종료 이벤트를 삽입할 수 있다.

[0087] 프로세서(200)는 연산이 다른 연산과 관련된 이벤트와 오버랩되는 경우, 오버랩이 시작되는 부분에 종료 이벤트를 삽입할 수 있다. 프로세서(200)는 연산이 다른 연산과 관련된 이벤트와 오버랩되지 않는 경우, 미싱 이벤트의 다음 이벤트로부터 제2 시간만큼 뺀 시간에 종료 이벤트를 삽입할 수 있다. 제2 시간은 연산의 종류와 하드웨어에 따라 상이할 수 있다. 예를 들어, 제2 시간은 10ns일 수 있다.

[0088] 프로세서(200)는 이벤트가 오버랩되는 경우, 오버랩된 부분의 동작은 타당(valid)하지 않은 것으로 판단할 수 있다. 프로세서(200)는 오버랩 되는 부분이 없도록 미싱 이벤트를 보완할 수 있다.

[0089] 시작 이벤트와 종료 이벤트가 둘 다 사라지거나, 3 개 이상의 이벤트가 사라진 경우, 프로세서(200)는 미싱 이벤트 이후에, 최초로 수신한 이벤트가 발생한 시간을 기준으로 제3 시간에 미리 결정된 인덱스(index)를 곱한 시간만큼 뺀 시간에 미싱 이벤트들을 삽입함으로써 보완을 수행할 수 있다.

[0090] 제3 시간은 연산의 종류와 하드웨어에 따라 상이할 수 있다. 예를 들어서, 제3 시간은 10ns일 수 있다.

[0092] 도 5는 도 1에 도시된 프로파일링 장치의 시각화 동작의 순서를 나타낸다.

[0093] 도 5를 참조하면, 프로세서(200)는 수신한 이벤트를 파싱(parsing)할 수 있다(510). 프로세서(200)는 이벤트 정보(예를 들어, 이벤트 파일)에 기록된 이벤트 패킷(event packet)을 확인함으로써 이벤트를 파싱을 수행할 수 있다. 이벤트 패킷은 타임스탬프(timestamp), 이벤트 ID(Identification) 및 이벤트의 종류를 포함할 수 있다. 이벤트 ID는 도 4에서 설명한 File, PU, RU를 포함할 수 있고, 이벤트의 종류는 시작 이벤트 및 종료 이벤트를 포함할 수 있다.

[0094] 프로세서(200)는 실행 순서와 이벤트가 매칭되는지 여부를 판단할 수 있다(520). 구체적으로, 프로세서(200)는 제어 프로그램에 포함된 실행 순서에 맞게 뉴럴 네트워크 연산의 시작 이벤트와 종료 이벤트가 수신되는지 여부를 판단함으로써 매칭 여부를 판단할 수 있다.

[0095] 프로세서(200)는 뉴럴 네트워크 연산의 실행 순서와 이벤트가 매칭되는 경우에 이벤트 로그를 출력할 수 있다(530). 실행 순서와 이벤트가 매칭되지 않는 경우, 프로세서(200)는 미싱 이벤트 로그를 출력할 수 있다(540).

[0096] 프로세서(200)는 이벤트 로그 및 미싱 이벤트 로그 출력을 통해 프로파일을 생성할 수 있다. 이벤트 로그 및 미싱 이벤트 로그 출력이 완료된 후, 프로세서(200)는 제어 프로그램을 종료할 수 있다(550). 제어 프로그램이 종료된 후, 프로세서(200)는 생성된 프로파일을 시각화할 수 있다.

[0098] 도 6은 도 1에 도시된 프로파일링 장치의 동작의 순서를 나타낸다.

[0099] 도 6을 참조하면, 수신기(100)는 뉴럴 네트워크 연산에 관련된 이벤트(event) 및 뉴럴 네트워크 연산을 수행하기 위한 제어 프로그램(control program)을 수신할 수 있다(610).

[0100] 프로세서(200)는 이벤트 및 제어 프로그램에 기초하여 미싱 이벤트(missing event)를 감지할 수 있다(630). 프로세서(200)는 이벤트가 상기 제어 프로그램에 포함된 실행 순서에 매칭되는지 여부를 판단할 수 있다. 프로세서(200)는 판단 결과에 기초하여 미싱 이벤트를 감지할 수 있다.

[0101] 프로세서(200)는 미싱 이벤트의 감지 결과에 기초하여 뉴럴 네트워크 연산에 대한 프로파일(profile)을 생성할

수 있다(650). 프로세서(200)는 미싱 이벤트의 종류를 결정할 수 있다. 프로세서(200)는 결정된 종류에 기초하여 미싱 이벤트를 보완함으로써 프로파일을 생성할 수 있다.

[0102] 프로세서(200)는 미싱 이벤트의 종류가 시작 이벤트인 경우, 미싱 이벤트의 다음 이벤트로부터 제1 시간만큼 뺀 시간에 시작 이벤트를 삽입(insert)할 수 있다.

[0103] 프로세서(200)는 미싱 이벤트의 종류가 종료 이벤트인 경우, 뉴럴 네트워크 연산이 다른 연산과 관련된 이벤트와 오버랩되는지 여부를 판단할 수 있다. 프로세서(200)는 판단 결과에 기초하여 종료 이벤트를 삽입할 수 있다.

[0104] 프로세서(200)는 연산이 다른 연산과 관련된 이벤트와 오버랩되는 경우, 오버랩이 시작되는 부분에 종료 이벤트를 삽입할 수 있다. 프로세서(200)는 연산이 다른 연산과 관련된 이벤트와 오버랩되지 않는 경우, 미싱 이벤트의 다음 이벤트로부터 제2 시간만큼 뺀 시간에 종료 이벤트를 삽입할 수 있다.

[0106] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다. 상기된 하드웨어 장치는 실시예의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

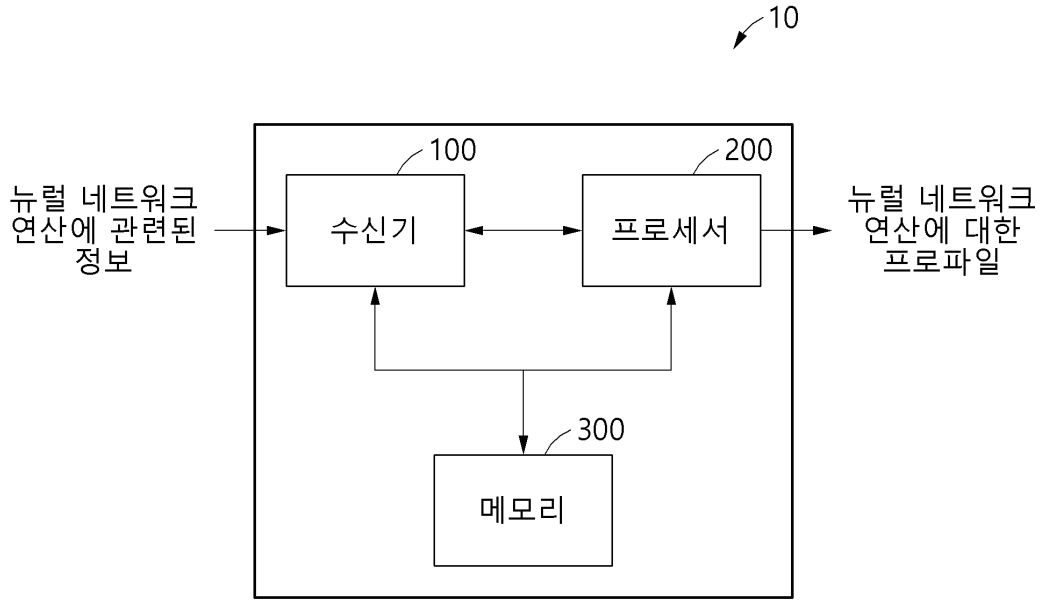
[0108] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상 장치(virtual equipment), 컴퓨터 저장 매체 또는 장치, 또는 전송되는 신호 파(signal wave)에 영구적으로, 또는 일시적으로 구체화(embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

[0110] 이상과 같이 실시예들이 비록 한정된 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기를 기초로 다양한 기술적 수정 및 변형을 적용할 수 있다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.

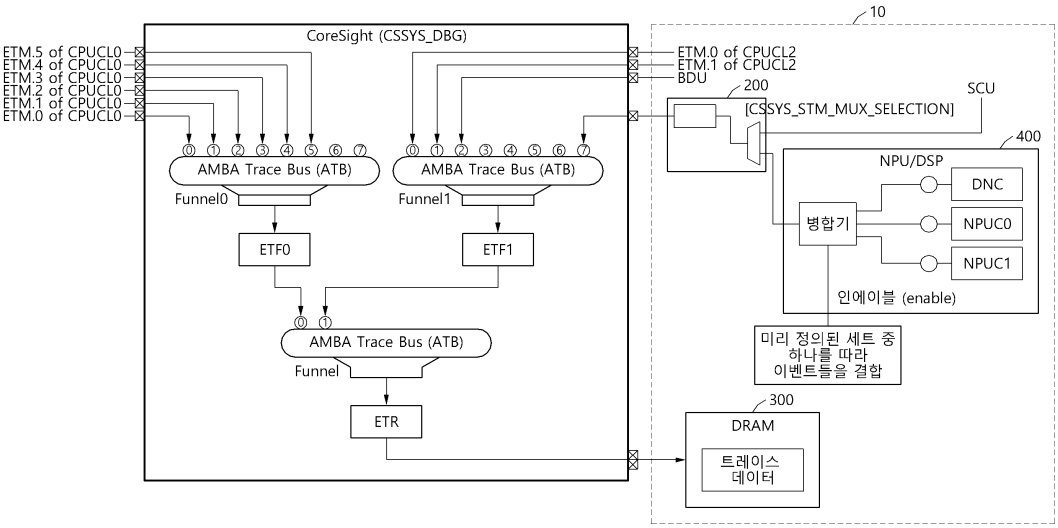
[0111] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 청구범위의 범위에 속한다.

도면

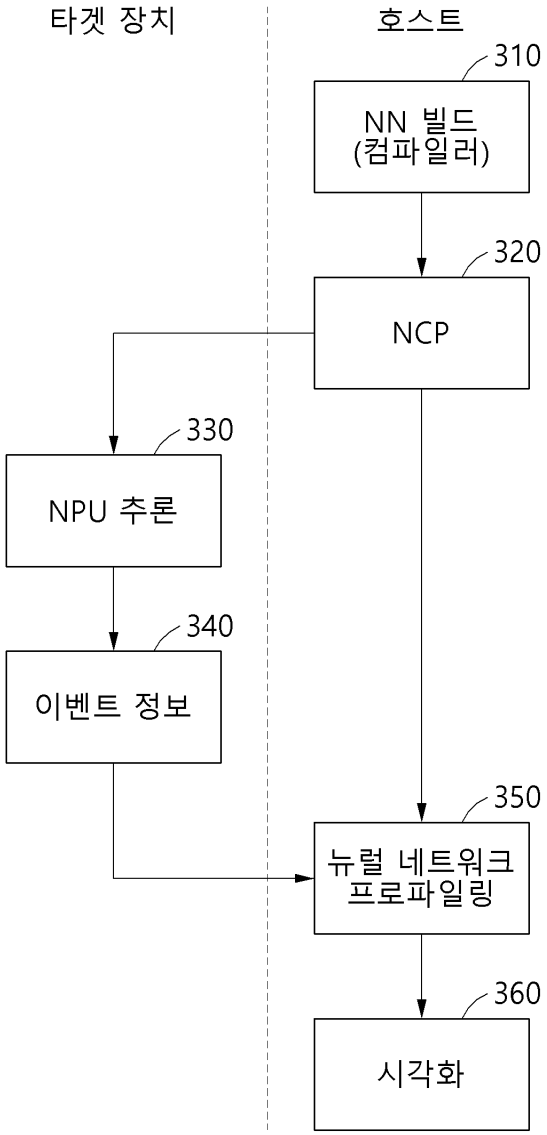
도면1


10
100
200
뉴럴 네트워크 연산에 관련된 정보
수신기
프로세서
뉴럴 네트워크 연산에 대한 프로파일
300
메모리

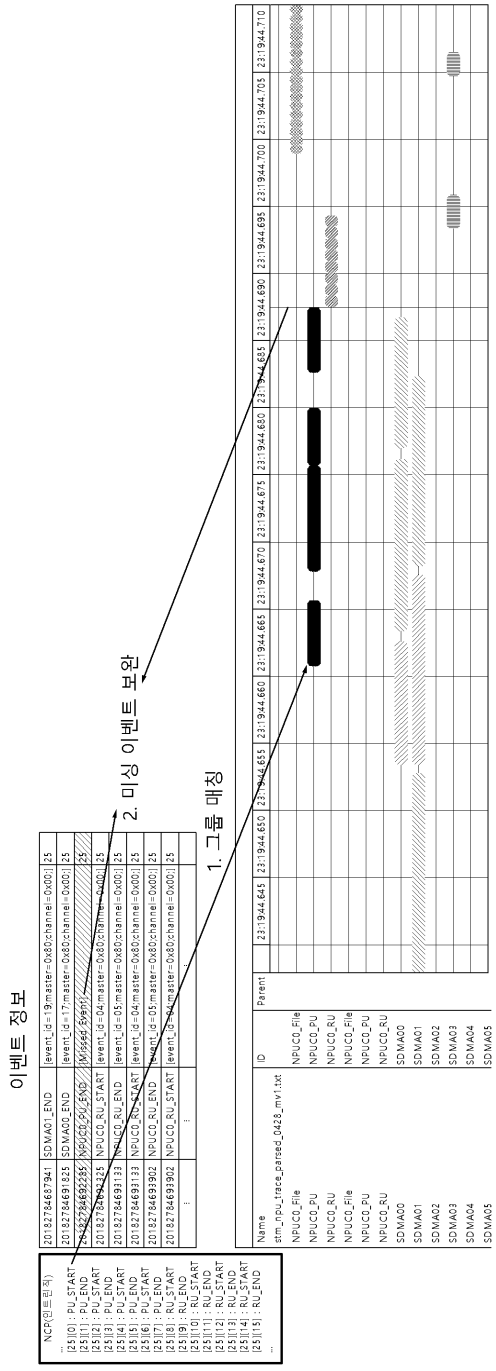
도면2


10
CoreSight (CSSYS_DBG)
ETM.5 of CPUCL0
ETM.4 of CPUCL0
ETM.3 of CPUCL0
ETM.2 of CPUCL0
ETM.1 of CPUCL0
ETM.0 of CPUCL0
ETM.0 of CPUCL2
ETM.1 of CPUCL2
BDU
200
[CSSYS_STM_MUX_SELECTION]
SCU
400
NPU/DSP
DNC
병합기
NPUC0
NPUC1
인에이블 (enable)
미리 정의된 세트 중 하나를 따라 이벤트들을 결합
AMBA Trace Bus (ATB)
Funnel0
AMBA Trace Bus (ATB)
Funnel1
ETF0
ETF1
AMBA Trace Bus (ATB)
Funnel
ETR
300
DRAM
트레이스 데이터

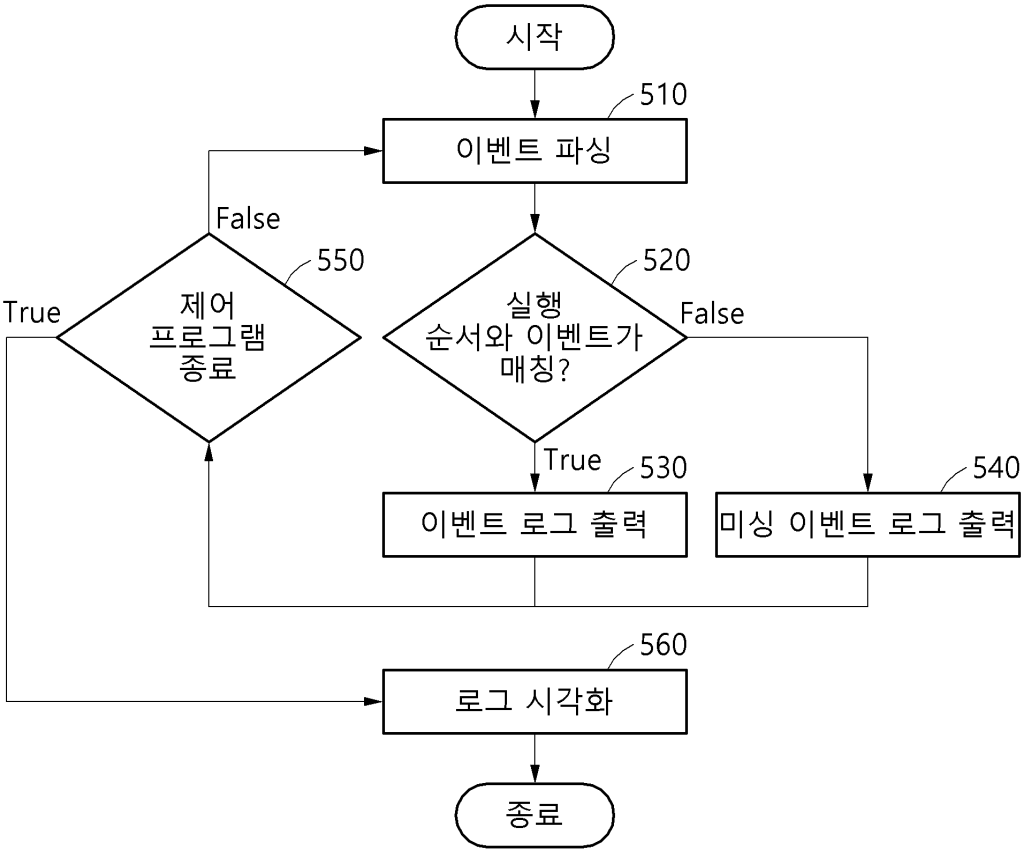
도면3


타겟 장치
호스트
310
NN 빌드
(컴파일러)
320
NCP
330
NPU 추론
340
이벤트 정보
350
뉴럴 네트워크
프로파일링
360
시각화

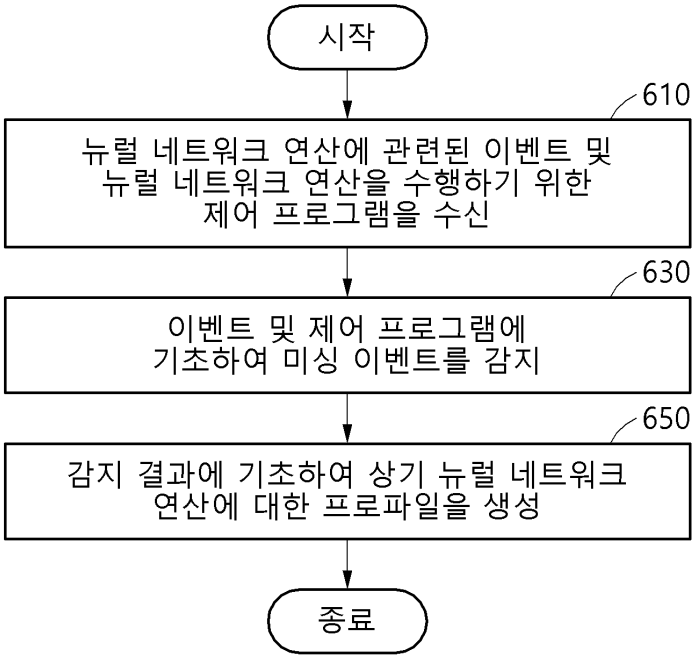
도면4



도면5


시작
510
이벤트 파싱
520
실행 순서와 이벤트가 매칭?
False
True
530
이벤트 로그 출력
540
미싱 이벤트 로그 출력
550
제어 프로그램 종료
False
True
560
로그 시각화
종료

도면6


시작
610
뉴럴 네트워크 연산에 관련된 이벤트 및 뉴럴 네트워크 연산을 수행하기 위한 제어 프로그램을 수신
630
이벤트 및 제어 프로그램에 기초하여 미싱 이벤트를 감지
650
감지 결과에 기초하여 상기 뉴럴 네트워크 연산에 대한 프로파일을 생성
종료