(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2023년04월06일
(11) 등록번호 10-2519467
(24) 등록일자 2023년04월04일

(51) 국제특허분류(Int. Cl.)
G06F 12/0811 (2016.01) *G06F 12/0804* (2016.01)
G06F 12/0888 (2016.01) *G06N 3/04* (2023.01)
G06N 3/063 (2023.01)
(52) CPC특허분류
G06F 12/0811 (2013.01)
G06F 12/0804 (2013.01)
(21) 출원번호 10-2019-7036813
(22) 출원일자(국제) 2019년06월27일
심사청구일자 2022년04월26일
(85) 번역문제출일자 2019년12월12일
(65) 공개번호 10-2021-0044669
(43) 공개일자 2021년04월23일
(86) 국제출원번호 PCT/CN2019/093144
(87) 국제공개번호 WO 2020/042739
국제공개일자 2020년03월05일
(30) 우선권주장
201810987293.5 2018년08월28일 중국(CN)
201810987343.X 2018년08월28일 중국(CN)
(56) 선행기술조사문헌
WO2018113239 A1
US20170221176 A1
US20160054922 A1

(73) 특허권자
캠브리콘 테크놀로지스 코퍼레이션 리미티드
중국 베이징 100191, 하이뎬 디스트릭트, 지춘 로드, 넘버 7, 즈쩐 빌딩, 블록 디, 16/에프, 룸 1601
(72) 발명자
리우 샤올리
중국 베이징 100190 하이뎬 디스트릭트 커쉐위안 사우스 로드 넘버 6 사이언티픽 리서치 빌딩 스위트 644
멩 샤오푸
중국 베이징 100190 하이뎬 디스트릭트 커쉐위안 사우스 로드 넘버 6 사이언티픽 리서치 빌딩 스위트 644
(74) 대리인
제일특허법인(유)

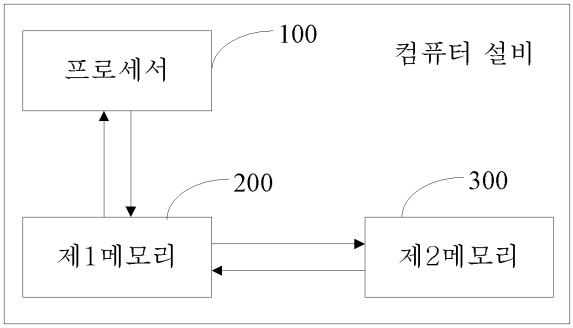
전체 청구항 수 : 총 10 항 심사관 : 안지현

(54) 발명의 명칭 **데이터 전처리 방법, 장치, 컴퓨터 설비 및 저장 매체**

(57) 요 약

본 개시는 데이터 전처리 방법, 장치, 컴퓨터 설비 및 저장 매체에 관한 것이며, 목표 연산 동작에 대응하는 목표 출력 데이터를 프로세서에 인접한 제 1 메모리에 저장하여, 목표 출력 데이터의 판독 횟수를 감소시킴으로써, 연산 과정에서의 I/O 판독 동작의 점유 시간을 단축할 수 있어, 프로세서의 속도 및 효율을 향상시킬 수 있다.

대 표 도 - 도1



(52) CPC특허분류
G06F 12/0888 (2013.01)
G06N 3/045 (2023.01)
G06N 3/063 (2013.01)
G06F 2212/1016 (2013.01)
G06F 2212/1044 (2013.01)
G06F 2212/454 (2013.01)
G06F 2212/604 (2013.01)

명 세 서

청구범위

청구항 1

데이터를 전처리하기 위한 방법으로서,

제 1 메모리의 가용 저장 용량 및 목표 연산 동작을 획득하는 단계;

상기 목표 연산 동작 및 상기 제 1 메모리의 가용 저장 용량에 따라, 상기 목표 연산 동작에 대응하는 목표 입력 데이터를 확정하는 단계 －상기 목표 입력 데이터는 상기 목표 연산 동작에 대응하는 입력 데이터의 일부 또는 전부임－;

상기 목표 연산 동작 및 상기 목표 입력 데이터에 따라, 상기 목표 연산 동작의 목표 출력 데이터를 확정하는 단계;

상기 목표 연산 동작의 목표 출력 데이터가 상기 목표 연산 동작 후의 기타 연산 동작의 입력 데이터인 경우, 상기 목표 연산 동작의 목표 출력 데이터를 상기 제 1 메모리에 저장하는 단계를 포함하고,

상기 제 1 메모리는 프로세서와 직접 데이터를 교환하기 위해 상기 프로세서에 인접하여 배치되는 것을 특징으로 하는,

방법.

청구항 2

제 1 항에 있어서,

상기 방법은, 상기 목표 연산 동작의 목표 출력 데이터가 상기 목표 연산 동작 후의 기타 연산 동작의 입력 데이터인 경우, 상기 목표 연산 동작의 목표 출력 데이터를 상기 제 1 메모리 및 제 2 메모리에 동기화 저장하는 단계를 더 포함하며;

상기 제 2 메모리는 상기 프로세서의 외부 메모리이며 상기 프로세서로부터 원격으로 배치되고, 상기 제 1 메모리의 저장 용량은 상기 제 2 메모리의 저장 용량보다 작은 것을 특징으로 하는,

방법.

청구항 3

제 1 항에 있어서,

상기 목표 연산 동작은 하나 이상의 연산 동작을 포함하고, 상기 연산 동작 각각은 서브 목표 입력 데이터에 대응되고;

상기의 목표 연산 동작 및 상기 제 1 메모리의 가용 저장 용량에 따라 상기 목표 연산 동작에 대응하는 목표 입력 데이터를 확정하는 단계는,

상기 제 1 메모리의 가용 저장 용량 및 처리하고자 하는 연산에서 각 연산 동작의 융합 속성에 따라, 융합 가능한 연산 동작의 수량을 확정하여 융합 수량의 역치를 얻는 단계;

수량이 선정된 상기 융합 가능한 연산 동작의 조합을 상기 목표 연산 동작으로 사용하고, 상기 선정된 수량은 상기 융합 수량 역치 이하인 단계;

상기 수량이 선정된 각 융합 가능한 연산 동작에 대응하는 서브 목표 입력 데이터를 상기 목표 연산 동작에 대응하는 목표 입력 데이터로 사용하는 단계를 더 포함하는 것을 특징으로 하는,

방법.

청구항 4

제 3 항에 있어서,

상기 처리하고자 하는 연산은 복수의 연산 계층을 포함하는 신경망 연산이며, 상기 연산 계층 각각은 하나의 상기 연산 동작을 나타내며;

상기 방법은,

상기 신경망 연산의 각 연산 계층의 연결 관계에 따라, 상기 연산 동작 각각의 융합 속성을 확정하는 단계를 더 포함하는 것을 특징으로 하는,

방법.

청구항 5

제 3 항에 있어서,

상기 방법은,

상기 목표 연산 동작에서 현재 연산 동작이 출력한 중간 계산 결과가 상기 목표 연산 동작 후의 기타 연산 동작의 입력 데이터인 경우, 또는 상기 현재 연산 동작이 출력한 중간 계산 결과가 기타 목표 연산 동작의 입력 데이터인 경우, 상기 현재 연산 동작이 출력한 중간 계산 결과는 제 1 메모리 상에 저장되거나, 혹은 상기 현재 연산 동작에 의해 출력된 중간 계산 결과는 제 1 메모리와 제 2 메모리에 동기화 저장되는 단계를 더 포함하는 것을 특징으로 하는,

방법.

청구항 6

제 1 항에 있어서,

상기 목표 연산 동작에 대응하는 입력 데이터는 복수의 입력 데이터 블록을 포함하고, 상기 목표 입력 데이터 각각은 하나 이상의 상기 입력 데이터 블록을 포함하며, 상기 목표 연산 동작에 대응하는 목표 입력 데이터의 수량은 하나 이상인 것을 특징으로 하는,

방법.

청구항 7

제 6 항에 있어서,

상기 목표 연산 동작은 하나 이상의 서브 목표 연산 동작을 포함하고, 상기 서브 목표 연산 동작 각각은 하나의 목표 입력 데이터를 대응하며; 상기 방법은,

상기 서브 목표 연산 동작 각각의 목표 입력 데이터의 데이터 용량 및 목표 출력 데이터의 데이터 용량에 따라, 상기 서브 목표 연산 동작 각각에 필요한 목표 저장 용량을 확정하는 단계;

상기 제 1 메모리의 가용 저장 용량 및 현재의 서브 목표 연산 동작에 필요한 목표 저장 용량에 따라, 상기 제 1 메모리의 잔여 저장 용량을 확정하는 단계;

상기 제 1 메모리의 잔여 저장 용량 및 상기 현재 서브 목표 연산 동작 이외의 기타 서브 목표 연산 동작에 필요한 목표 저장 용량에 따라, 상기 서브 목표 연산 동작의 수량을 확정하는 단계를 더 포함하는 것을 특징으로 하는,

방법.

청구항 8

제 1 항에 있어서,

상기 방법은,

상기 목표 연산 동작 후의 기타 연산 동작이 상기 목표 연산 동작의 목표 입력 데이터를 사용할 필요가 없는 경우, 상기 목표 연산 동작이 완료된 후, 상기 목표 연산 동작에 대응하는 목표 입력 데이터의 저장 주소의 일부 또는 전부를 상기 목표 연산 동작의 목표 출력 데이터에 분배하는 단계를 더 포함하는 것을 특징으로 하는,

방법.

청구항 9

제 1 메모리, 제 2 메모리 및 프로세서를 포함하고, 상기 제 1 메모리는 상기 프로세서와 직접 데이터를 교환하기 위해 상기 프로세서에 근접하여 배치되고, 상기 제 1 메모리와 상기 제 2 메모리는 데이터를 리드-라이트 할 수 있으며; 상기 제 1 메모리 또는 상기 제 2 메모리에 컴퓨터 프로그램이 내장된 컴퓨터 설비로서,

상기 컴퓨터 프로그램은 상기 프로세서에 의해 실행될 경우, 제1항 내지 제8항 중 어느 한 항에 따른 데이터를 전처리하기 위한 방법의 단계를 수행하는 것을 특징으로 하는,

컴퓨터 설비.

청구항 10

컴퓨터 프로그램이 내장된 컴퓨터 판독 가능 저장 매체로서,

상기 컴퓨터 프로그램은 프로세서에 의해 실행될 경우, 제1항 내지 제8항 중 어느 한 항에 따른 데이터를 전처리하기 위한 방법의 단계를 수행하는 것을 특징으로 하는,

컴퓨터 판독 가능 저장 매체.

청구항 11

삭제

청구항 12

삭제

청구항 13

삭제

청구항 14

삭제

청구항 15

삭제

청구항 16

삭제

청구항 17

삭제

청구항 18

삭제

청구항 19

삭제

청구항 20

삭제

발명의 설명

기 술 분 야

[0001] 본 개시는 2018년 8월 28일자, 중국 특허 출원 제2018109872935호, 명칭 "데이터 전처리 방법, 장치, 컴퓨터 설비 및 저장 매체" 2018년 8월 28일자, 중국 특허 출원 제 201810987343X호, 명칭 "데이터 전처리 방법, 장치, 컴퓨터 설비 및 저장 매체"에 기초한 우선권의 이익을 주장하며, 해당 중국 특허 출원의 문헌에 개시된 모든 내용은 본 명세서의 일부로서 포함한다.

[0002] 본 개시는 컴퓨터 기술 분야에 관한 것으로서, 특히 데이터 전처리 방법, 장치, 컴퓨터 설비 및 저장 매체에 관한 것이다.

배 경 기 술

[0003] 데이터 량이 폭발적으로 증가함에 따라 기계 학습과 같은 인공 지능 알고리즘이 점점 더 많이 응용되고 있다. 기기는 많은 양의 데이터를 분석하여 학습하므로, 기계 학습 등과 같은 빅데이터 연산은 메모리 액세스 등에 대한 수요가 급격히 증가하고 있다.

[0004] 메모리 액세스 등의 수요를 충족시키기 위해, 현재는 일반적으로 캐시 메모리, 메인 메모리 및 외장 메모리를 사용하는 멀티 레벨 메모리 아키텍처가 사용된다. 여기서, 캐시 메모리(Cache) , 메인 메모리 및 외장 메모리의 액세스 속도는 순차적으로 감소하고, 저장 용량은 순차적으로 증가한다. 하지만, 컴퓨터 설비에서 I/O의 대역폭은 종종 많은 양의 데이터 수요를 충족시킬 수 없기 때문에, 프로세서가 기계 학습 연산을 실행하는 과정 중에, 캐시 메모리와 메인 메모리 사이, 및/또는 메인 메모리와 외장 메모리 사이에서 데이터 판독 동작은 빈번하게 수행될 필요가 있다. 예를 들어, 프로세서가 연산을 실행하는 과정에서, 프로세서는 먼저 외장 메모리로부터 입력 데이터를 판독하고, 연산이 종료된 후, 프로세서는 연산 결과를 외장 메모리에 저장하고, 다음 연산에 필요할 입력 데이터를 외장 메모리로부터 계속 판독한다. I/O 대역폭의 제한으로 인해 하나의 연산 과정에는 적어도 두 번의 I/O 리드-라이트 작업이 필요하며, 빈번한 I/O 리드-라이트 작업은 많은 시간을 차지하여 프로세서의 처리 효율이 낮춘다.

발명의 내용

[0005] 이를 감안하여, 상기 기술적 문제를 해결하기 위한, 연산 과정에서 I/O 리드-라이트 동작의 횟수를 감소시키고 프로세서의 처리 효율을 향상시킬 수 있는 데이터 전처리 방법, 장치, 컴퓨터 설비 및 저장 매체를 제공할 필요가 있다.

[0006] 데이터 전처리 방법으로서, 상기 방법은,

[0007] 제 1 메모리의 가용 저장 용량 및 목표 연산 동작을 획득하는 단계;

[0008] 목표 연산 동작 및 제 1 메모리의 가용 저장 용량에 따라, 상기 목표 연산 동작에 대응하는 목표 입력 데이터를 확정하는 단계; 여기서, 상기 목표 입력 데이터는 상기 목표 연산 동작에 대응하는 모든 입력 데이터의 일부 또는 전부이며;

[0009] 상기 목표 연산 동작 및 상기 목표 입력 데이터에 따라, 상기 목표 연산 동작의 목표 출력 데이터를 확정하는

단계;

[0010] 상기 목표 연산 동작의 목표 출력 데이터가 상기 목표 연산 동작 후의 기타 연산 동작의 입력 데이터인 경우, 상기 목표 연산 동작의 목표 출력 데이터를 상기 제 1 메모리에 저장하는 단계를 포함하며, 여기서, 상기 제 1 메모리는 프로세서에 인접하여 배치된다.

[0011] 데이터 전처리 장치로서, 상기 장치는,

[0012] 제 1 메모리의 가용 저장 용량 및 목표 연산 동작을 획득하기 위한 획득 모듈;

[0013] 상기 목표 연산 동작 및 상기 제 1 메모리의 가용 저장 용량에 따라, 상기 목표 연산 동작에 대응하는 목표 입력 데이터를 확정하기 위한 입력 확정 모듈;

[0014] 상기 목표 연산 동작 및 상기 목표 입력 데이터에 따라, 상기 목표 연산 동작의 목표 출력 데이터를 확정하기 위한 출력 확정 모듈;

[0015] 저장 할당 모듈 상기 목표 연산 동작의 목표 출력 데이터가 상기 목표 연산 동작 후의 기타 연산 동작의 입력 데이터인 경우, 상기 목표 연산 동작의 목표 출력 데이터를 상기 제 1 메모리에 저장하기 위한 저장 할당 모듈을 포함하고, 여기서, 상기 제 1 메모리는 프로세서에 인접하여 배치된다.

[0016] 제 1 메모리, 제 2 메모리 및 프로세서를 포함하며, 상기 제 1 메모리는 프로세서에 인접하여 배치되고, 상기 제 1 메모리와 상기 제 2 메모리는 데이터를 리드-라이트할 수 있으며; 상기 제 2 메모리에는 컴퓨터 프로그램이 내장된 컴퓨터 설비로서, 상기 컴퓨터 프로그램은 상기 프로세서에 의해 실행될 경우, 상기 데이터 전처리 방법의 단계를 수행하는 것을 특징으로 한다.

[0017] 컴퓨터 프로그램이 내장된 컴퓨터 판독 가능 저장 매체로서, 상기 컴퓨터 프로그램은 프로세서에 의해 실행될 경우, 상기 데이터 전처리 방법의 단계를 수행한다.

[0018] 상기 데이터 전처리 방법, 장치, 컴퓨터 설비 및 저장 매체에 있어서, 상기 목표 연산 동작의 목표 출력 데이터가 그 다음의 기타 연산 동작의 입력 데이터인 경우, 목표 연산 동작에 대응하는 목표 출력 데이터를 프로세서에 인접하여 위치한 제 1 메모리에 저장할 수 있기에, 목표 출력 데이터의 판독 횟수를 감소시킴으로써, 연산 과정에서의 I/O 판독 동작의 점유 시간이 감소되고, 이에 의해 프로세서의 속도 및 효율이 개선된다.데이터 전처리 방법으로서, 상기 방법은,

[0019] 메인 메모리의 가용 저장 용량, 슬레이브 메모리에서 가용 저장 용량 및 목표 연산 동작을 획득하는 단계;

[0020] 상기 메인 메모리의 가용 저장 용량, 상기 슬레이브 메모리의 가용 저장 용량 및 상기 목표 연산 동작에 따라, 상기 목표 연산 동작에 대응하는 목표 입력 데이터를 확정하는 단계;

[0021] 상기 목표 연산 동작 및 상기 목표 입력 데이터에 따라, 상기 목표 연산 동작에 대응하는 목표 출력 데이터를 확정하는 단계;

[0022] 상기 목표 연산 동작의 목표 출력 데이터가 상기 목표 연산 동작 후의 기타 연산 동작의 입력 데이터인 경우, 상기 목표 출력 데이터는 상기 메인 메모리에 대응하게 저장되는 단계를 포함한다.

[0023] 일 실시예에서, 상기 메인 메모리의 가용 저장 용량, 상기 슬레이브 메모리의 가용 저장 용량 및 상기 목표 연산 동작에 따라, 상기 목표 연산 동작에 대응하는 목표 입력 데이터를 확정하는 상기 단계는,

[0024] 상기 메인 메모리의 가용 저장 용량을 상기 슬레이브 메모리의 가용 저장 용량 각각에 비교하여, 최소의 가용 저장 용량을 제 1 메모리의 가용 저장 용량으로 사용하는 단계;

[0025] 상기 제 1 메모리의 가용 저장 용량 및 상기 목표 연산 동작에 따라, 상기 목표 연산 동작에 대응하는 목표 입력 데이터를 확정하는 단계를 더 포함한다.

[0026] 일 실시예에서, 상기 목표 연산 동작은 하나 이상의 연산 동작을 포함하며, 모든 상기 연산 동작은 대응하는 서브 목표 입력 데이터가 있으며; 목표 연산 동작 및 제 1 메모리의 가용 저장 용량에 따라 상기 목표 연산 동작에 대응하는 목표 입력 데이터를 확정하는 상기 단계는,

[0027] 상기 제 1 메모리의 가용 저장 용량 및 처리하고자 하는 연산에서 각 연산 동작의 융합 속성에 따라, 융합 가능한 연산 동작의 수량을 확정하고, 융합 수량 역치를 얻는 단계;

[0028] 수량이 선정된 상기 융합 가능한 연산 동작의 조합을 상기 목표 연산 동작으로 사용하고, 상기 선정된 수량은

융합 수량 역치이하인 단계;

[0029] 상기 수량이 선정된 융합 가능한 연산 동작 각각에 대응하는 서브 목표 입력 데이터를 상기 목표 연산 동작에 대응하는 목표 입력 데이터로 사용하는 단계를 더 포함한다.

[0030] 일 실시예에서, 상기 처리하고자 하는 연산은 복수의 연산 계층을 포함하는 신경망 연산이며, 상기 연산 계층 각각은 하나의 상기 연산 동작을 나타내며, 상기 방법은,

[0031] 상기 신경망 연산의 각 연산 계층의 연결 관계에 따라, 상기 연산 동작 각각의 융합 속성을 확정하는 단계를 더 포함한다.

[0032] 일 실시예에서, 상기 목표 연산 동작에 대응하는 입력 데이터는 복수의 입력 데이터 블록을 포함하고, 상기 목표 입력 데이터 각각은 하나 이상의 상기 입력 데이터 블록을 포함하며, 상기 목표 연산 동작에 대응하는 목표 입력 데이터의 수량은 하나 이상이다.

[0033] 일 실시예에서, 상기 목표 연산 동작은 하나 이상의 서브 목표 연산 동작을 포함하고, 상기 서브 목표 연산 동작 각각은 하나의 목표 입력 데이터에 대응되며, 상기 방법은,

[0034] 상기 서브 목표 연산 동작의 목표 입력 데이터의 데이터 용량 및 목표 출력 데이터의 데이터 용량에 각각 따라, 상기 서브 목표 연산 동작 각각에 필요한 목표 저장 용량을 확정하는 단계;

[0035] 상기 제 1 메모리의 가용 저장 용량 및 현재의 서브 목표 연산 동작에 필요한 목표 저장 용량에 따라, 상기 제 1 메모리의 잔여 저장 용량을 확정하는 단계;

[0036] 상기 제 1 메모리의 잔여 저장 용량 및 상기 현재 서브 목표 연산 동작 이외의 기타 서브 목표 연산 동작에 필요한 목표 저장 용량에 따라, 상기 서브 목표 연산 동작의 수량을 확정하는 단계를 더 포함한다.

[0037] 일 실시예에서, 상기 목표 입력 데이터는 제 1 목표 입력 데이터 및 제 2 목표 입력 데이터를 포함하고, 상기 방법은,

[0038] 미리 설정된 연산 할당 규칙에 따라, 상기 메인 메모리에 대응하는 제 1 목표 입력 데이터 및 상기 슬레이브 메모리에 대응하는 각각 제 2 목표 입력 데이터를 확정하는 단계를 더 포함한다.

[0039] 일 실시예에서, 상기 방법은:

[0040] 상기 메인 메모리의 가용 저장 용량 및 상기 제 1 목표 입력 데이터의 데이터 용량에 따라, 상기 메인 메모리 상에 있는 상기 제 1 목표 입력 데이터의 저장 주소를 확정하는 단계;

[0041] 상기 슬레이브 메모리 각각의 가용 저장 용량 및 그에 대응하는 상기 제 2 목표 입력 데이터의 데이터 용량에 따라, 상기 슬레이브 메모리 상에 있는 상기 제 2 목표 입력 데이터 각각의 저장 주소를 확정하는 단계를 더 포함한다.

[0042] 일 실시예에서, 상기 목표 출력 데이터는 제 1 목표 출력 데이터 및 제 2 목표 출력 데이터를 포함하고; 상기 목표 연산 동작 및 상기 목표 입력 데이터에 따라, 상기 목표 연산 동작에 대응하는 목표 출력 데이터를 확정하는 상기 단계은,

[0043] 상기 목표 연산 동작 및 상기 제 1 목표 입력 데이터에 따라, 상기 제 1 목표 출력 데이터 및 상기 제 1 목표 출력 데이터가 상기 메인 메모리 상에서 저장된 주소를 확정되는 단계;

[0044] 상기 목표 연산 동작 및 상기 제 2 목표 입력 데이터 각각에 따라, 상기 제 2 목표 출력 데이터 및 상기 제 2 목표 출력 데이터 각각에 해당되는 슬레이브 메모리 상에서 저장된 주소를 확정하는 단계;

[0045] 상기 제 2 목표 출력 데이터 각각에 따라, 상기 메인 메모리 상에 있는 상기 제 2 목표 출력 데이터 각각의 저장 주소를 확정하는 단계를 더 포함한다.

[0046] 일 실시예에서, 상기 방법은,

[0047] 상기 슬레이브 프로세싱 회로 상에서 수행되는 기타 목표 연산 동작에 상기 제 2 목표 출력 데이터를 사용할 수요가 있을 경우, 상기 제 2 목표 출력 데이터를 상기 슬레이브 프로세싱 회로에 대응하는 슬레이브 메모리에 저장하는 단계를 더 포함한다.

[0048] 일 실시예에서, 상기 방법은,

[0049] 상기 목표 연산 동작의 목표 출력 데이터가 상기 목표 연산 동작 후의 기타 연산 동작의 입력 데이터인 경우, 상기 목표 출력 데이터는 대응하여 상기 메인 메모리와 상기 제 2 메모리 상에 저장되는 단계를 더 포함한다.

[0050] 데이터 전처리 장치로서, 상기 장치는:

[0051] 메인 메모리의 가용 저장 용량, 슬레이브 메모리의 가용 저장 용량 및 목표 연산 동작을 획득하기 위한 획득 모듈;

[0052] ,상기 메인 메모리의 가용 저장 용량, 상기 슬레이브 메모리의 가용 저장 용량 및 상기 목표 연산 동작에 따라, 상기 목표 연산 동작에 대응하는 목표 입력 데이터를 확정하기 위한 입력 확정 모듈;

[0053] 상기 목표 연산 동작 및 상기 목표 입력 데이터에 따라, 상기 목표 연산 동작에 대응하는 목표 출력 데이터를 확정하기 위한 출력 확정 모듈;

[0054] 저장 할당 모듈상기 목표 연산 동작의 목표 출력 데이터가 상기 목표 연산 동작 후의 기타 연산 동작의 입력 데이터인 경우, 상기 목표 출력 데이터는 메인 메모리에 대응하여 저장하기 위한 저장 할당 모듈를 포함한다.

[0055] 일 실시예에서, 상기 데이터 전처리 장치는 저장 용량 확정 모듈을 더 포함한다. 상기 저장 용량 확정 모듈은, 상기 메인 메모리의 가용 저장 용량과 상기 슬레이브 메모리 각각의 가용 저장 용량를 비교하여, 최소의 가용 저장 용량을 제 1 메모리의 가용 저장 용량으로 사용하도록 구성되며;

[0056] 입력 확정 모듈은, 구체적으로 상기 제 1 메모리의 가용 저장 용량 및 목표 연산 동작에 따라, 목표 연산 동작에 대응하는 목표 입력 데이터를 확정하도록 구성된다.

[0057] 일 실시예에서, 상기 목표 연산 동작은 하나 이상의 연산 동작을 포함하고, 상기 연산 동작 각각은 서브 목표 입력 데이터에 대응하고; 상기 입력 확정 모듈은 또한,

[0058] 상기 제 1 메모리의 가용 저장 용량 및 상기 처리하고자 하는 연산에서 각 연산 동작의 융합 속성에 따라, 융합 가능한 연산 동작의 수량을 확정하여 융합 수량 역치를 얻기 위한 융합 확정 유닛;

[0059] 수량이 선정된 상기 융합 가능한 연산 동작의 조합을 상기 목표 연산 동작으로 사용하고, 상기 선정된 수량은 상기 융합 수량 역치 이하이며, 상기 수량이 선정된 융합 가능한 연산 동작 각각 에 대응하는 서브 목표 입력 데이터를 상기 목표 연산 동작에 대응하는 목표 입력 데이터로 사용하기 위한 입력 확정 유닛을 더 포함한다.

[0060] 일 실시예에서, 상기 처리하고자 하는 연산은 복수의 연산 계층을 포함하는 신경망 연산이며, 상기 연산 계층 각각은 하나의 상기 연산 동작을 나타내며; 또한, 상기 융합 확정 유닛은 상기 신경망 연산의 각 연산 계층의 연결 관계에 따라 상기 각 연산 동작의 융합 속성을 확정하도록 구성된다.

[0061] 일 실시예에서, 상기 목표 연산 동작은 하나 이상의 서브 목표 연산 동작을 포함하고, 상기 서브 목표 연산 동작 각각은 하나의 상기 목표 입력 데이터에 대응하며; 여기서, 상기 목표 연산 동작에 대응하는 모든 입력 데이터는 복수의 입력 데이터 블록을 포함하고, 상기 목표 입력 데이터 각각은 하나 이상의 상기 입력 데이터 블록을 포함하고, 상기 목표 연산 동작에 대응하는 목표 입력 데이터의 수량은 하나 이상이며, 상기 입력 확정 모듈은 또한,

[0062] 상기 서브 목표 연산 동작 각각의 목표 입력 데이터의 데이터 용량 및 목표 출력 데이터의 데이터 용량에 각각 따라, 상기 서브 목표 연산 동작 각각에 필요한 목표 저장 용량을 확정하며;

[0063] 상기 제 1 메모리의 가용 저장 용량 및 현재의 서브 목표 연산 동작에 필요한 목표 저장 용량에 따라 상기 제 1 메모리의 잔여 저장 용량을 확정하며;

[0064] 상기 제 1 메모리의 잔여 저장 용량 및 상기 현재의 서브 목표 연산 동작 이외의 기타 서브 목표 연산 동작에 필요한 목표 저장 용량에 따라, 상기 서브 목표 연산 동작의 수량을 확정하도록 구성된다.

[0065] 일 실시예에서, 상기 목표 입력 데이터는 제 1 목표 입력 데이터 및 제 2 목표 입력 데이터를 포함하며;

[0066] 상기 입력 확정 모듈은 또한, 미리 설정된 연산 할당 규칙에 따라, 상기 메인 메모리에 대응하는 제 1 목표 입력 데이터 및 상기 슬레이브 메모리 각각에 대응하는 제 2 목표 입력 데이터를 확정하도록 구성되며;

[0067] 상기 저장 할당 모듈은 또한, 상기 메인 메모리의 가용 저장 용량 및 상기 제 1 목표 입력 데이터의 데이터 용량에 따라, 상기 메인 메모리 상에 있는 상기 제 1 목표 입력 데이터의 저장 주소를 확정하며; 상기 슬레이브 메모리 각각의 가용 저장 용량 및 그에 대응하는 상기 제 2 목표 입력 데이터의 데이터 용량에 각각 따라, 상기

슬레이브 메모리상에 있는 상기 제 2 목표 입력 데이터 각각의 저장 주소를 확정하도록 구성된다.

[0068] 일 실시예에서, 상기 목표 출력 데이터는 제 1 목표 출력 데이터 및 제 2 목표 출력 데이터를 포함하고; 상기 출력 확정 모듈은 또한,

[0069] 상기 목표 연산 동작 및 상기 제 1 목표 입력 데이터에 따라, 상기 제 1 목표 출력 데이터 및 상기 제 1 목표 출력 데이터가 상기 메인 메모리 상에서 저장된 주소를 확정하며;

[0070] 상기 목표 연산 동작 및 상기 제 2 목표 입력 데이터 각각에 따라, 상기 제 2 목표 출력 데이터 및 상기 제 2 목표 출력 데이터 각각은 해당 슬레이브 메모리 상에서 저장된 주소를 확정하며;

[0071] 상기 제 2 목표 출력 데이터 각각에 따라, 상기 메인 메모리 상에 있는 상기 제 2 목표 출력 데이터 각각의 저장 주소를 확정하도록 구성된다.

[0072] 일 실시예에서, 상기 저장 할당 모듈은 또한, 상기 슬레이브 프로세싱 회로 상에서 수행되는 기타 목표 연산 동작에 상기 제 2 목표 출력 데이터를 사용할 수요가 있을 경우, 상기 제 2 목표 출력 데이터를 상기 슬레이브 프로세싱 회로에 대응하는 슬레이브 메모리에 저장하도록 구성된다.

[0073] 컴퓨터 설비로서,

[0074] 서로 연결된 제어기 유닛과 연산 유닛을 포함하고, 상기 연산 유닛은 하나의 메인 프로세싱 회로 및 복수의 슬레이브 프로세싱 회로를 포함하는 프로세서;

[0075] 메인 메모리 및 복수의 슬레이브 메모리를 포함하고, 상기 메인 메모리는 상기 메인 프로세서에 인접하여 설치되고, 복수의 슬레이브 메모리는 복수의 상기 슬레이브 프로세싱 회로에 대응하여 설치되고, 상기 슬레이브 프로세서 각각은 대응한 상기 슬레이브 프로세싱 회로에 각각 인접하여 배치되는 복수의 제 1 메모리;

[0076] 제 2 메모리를 포함하며; 상기 제 1 메모리와 상기 제 2 메모리는 데이터를 리드-라이트 할 수 있으며;

[0077] 여기서, 상기 제 1 메모리 또는 제 2 메모리에 컴퓨터 프로그램이 저장되어 있고, 상기 프로세서가 상기 컴퓨터 프로그램을 실행할 때, 본 개시의 실시예 중 방법의 단계를 수행한다.

[0078] 컴퓨터 판독 가능 저장 매체에 있어서, 위에 저장된 컴퓨터 프로그램이 프로세서에 의해 실행될 때, 본 개시의 실시예 중 방법의 단계를 수행한다.

[0079] 상기 데이터 전처리 방법, 장치, 컴퓨터 설비 및 저장 매체에서, 상기 목표 연산 동작의 목표 출력 데이터가 그 다음의 기타 연산 동작의 입력 데이터일 경우, 목표 연산 동작에 대응하는 목표 출력 데이터를 메인 메모리에 저장할 수 있고, 메인 메모리와 제 2 메모리 사이의 데이터 상호작용을 감소시키고 목표 출력 데이터의 판독 횟수를 감소시킴으로써, 연산 과정에서의 I/O 판독 동작의 점유 시간이 감소되고, 이에 의해 프로세서의 속도 및 효율을 향상시킨다. 또한, 상기 데이터 전처리 방법은 메인 메모리와 슬레이브 메모리 사이의 데이터 상호작용을 감소시켜, 연산 과정에서의 I/O 판독 동작의 점유 시간이 한층 더 감소하여 프로세서의 속도 및 효율을 향상시킨다.

도면의 간단한 설명

[0080] 첨부된 도면은 명세서의 일부로 명세서에 속하며, 본 개시에 부합하는 실시예를 명시하였으며, 명세서와 함께 본 개시의 원리를 해석하기 위한 것이다.

도 1은 일 실시예에 따른 컴퓨터 설비의 구조를 제시하는 도면이다;

도 2는 일 실시예에 따른 컴퓨터 설비의 프로세서의 구조를 제시하는 도면이다;

도 3은 일 실시예에 따른 컴퓨터 설비의 프로세서의 구조를 제시하는 도면이다;

도 4는 일 실시예에 따른 컴퓨터 설비의 프로세서의 구조를 제시하는 도면이다;

도 5는 일 실시예에 따른 데이터 전처리 방법의 흐름을 제시하는 도면이다;

도 6은 도 5에 따른 목표 입력 데이터를 확정하는 단계의 일 실시예의 흐름을 제시하는 도면이다;

도7은 도 5에 따른 데이터 전처리 방법에서의 목표 연산 동작의 수량을 확정하는 일 실시예의 흐름을 제시하는 도면이다;

도 8은 일 실시예에 따른 처리하고자 하는 연산을 제시하는 도면이다;

도 9는 다른 실시예에 따른 처리하고자 하는 연산을 제시하는 도면이다;

도 10은 다른 실시예에 따른 데이터 전처리 방법의 흐름을 제시하는 도면이다;

도 11은 도 10 에 따른 목표 입력 데이터를 확정하는 단계의 일 실시예의 흐름을 제시하는 도면이다;

도 12는 일 실시예에 따른 데이터 전처리 장치의 구조 블록을 제시하는 도면이다;

도 13은 일 실시예에 따른 데이터 전처리 장치의 구조 블록을 제시하는 도면이다;

도 14는 다른 실시예에 따른 데이터 전처리 장치의 구조 블록을 제시하는 도면이다.

발명을 실시하기 위한 구체적인 내용

[0081] 본 개시의 목적, 기술 방안 및 장점을 보다 명확하게 하기 위해, 이하에서 첨부 도면 및 실시예를 결합하여 본 개시에 대해 보다 상세하게 설명한다. 본 명세서에 설명된 구체적인 실시예는 단지 본 개시를 설명하기 위한 것이며, 본 개시를 제한하는 것은 아닌 것으로 이해해야 한다. 본 개시의 청구 범위, 명세서 및 도면에서, 용어 "제1", "제2", "제3" 및 "제4"와 같은 용어는 다양한 개체들을 구별하기 위해 사용되는 것이고 특정 순서를 묘사하기 위해 사용되는 것은 아님을 이해해야 한다. 본 개시의 명세서와 청구 범위에서 사용되는 용어 "포함" 및 "함유"는 특징, 전체, 단계, 동작, 원소 및/또는 구성 요소의 존재를 나타내며, 하나 또는 복수의 기타 특징, 전체, 단계, 동작, 원소, 구성 요소 및/또는 그 집합의 존재 혹은 추가를 배제하지 않는다. 또한, 본 개시의 명세서에서 사용되는 용어는 단지 특정 실시예를 설명하기 위한 것이며, 본 개시를 제한하는 의도가 아님을 이해해야 한다. 본 개시의 명세서와 청구 범위에서 사용된 것과 같이, 앞뒤 문장에서 명확하게 기타 경우를 지적하지 않았다면, 단수 형태의 "일", "하나" 및 "이"는 복수 형태도 포함한다. 또한, 본 개시의 명세서와 청구 범위에서 사용된 용어 "및/또는"은 나열된 관련 항목 중 하나 또는 복수의 임의 조합 및 모든 가능한 조합을 가리키며, 이런 조합들을 포함한다는 것을 이해해야 한다. 본 개시의 명세서와 청구 범위에서 사용된 것과 같이, 용어 "만약"은 앞뒤 문맥에 따라 "…경우" 또는 "…하면" 또는 "확정에 대한 응답으로…" 또는 "검출에 대한 응답으로…"로 해석될 수 있다. 마찬가지로, 문구 "확정된다면" 또는 "[기술된 조건 또는 사건이] 검출된다면"은 앞뒤 문맥에 따라 "일단 확정되면" 또는 "확정에 대한 응답으로…" 또는 "[기술된 조건 또는 사건이] 일단 검출되면" 또는 "[기술된 조건 또는 사건이] 검출에 대한 응답으로…"로 해석될 수 있다. 도 1을 참조하면, 본 개시 실시예의 컴퓨터 설비는 프로세서(100), 제 1 메모리(200) 및 제 2 메모리(300)을 포함할 수 있다. 여기서, 제 1 메모리(200)는 프로세서(100)에 인접하여 배치될 수 있고, 프로세서(100)는 제 1 메모리(200)와 직접 데이터를 교환할 수 있고, 즉 프로세서(100)는 제 1 메모리(200)로부터 입력 데이터를 직접적으로 판독할 수 있으며, 상기 입력 데이터에 의해 얻어진 출력 데이터는 상기 제 1 메모리(200)에 기록될 수 있다. 상기 제 1 메모리(200)는 제 2 메모리(300)와 데이터 상호작용을 직접적으로 수행할 수 있고, 즉 상기 제 1 메모리(200)는 제 2 메모리(300)로부터 데이터를 직접 판독할 수도 있고, 상기 제 2 메모리(300)에 데이터를 기록할 수도 있다. 또한, 상기 제 1 메모리(200)의 액세스 속도는 제 2 메모리(300)의 액세스 속도보다 빠르고, 상기 제 1 메모리(200)의 저장 용량은 제 2 메모리(300)의 저장 용량보다 작다.

[0082] 선택적으로, 상기 컴퓨터 설비는 휴대 전화 또는 태블릿 컴퓨터 등과 같은 모바일 단말기, 또는 데스크탑 컴퓨터, 보드, 클라우드 서버 등과 같은 단말기 일 수 있다. 물론, 상기 컴퓨터 설비는 클라우드 서버 및 휴대 전화 또는 컴퓨터 등 단말기에 의해 형성된 컴퓨터 시스템일 수도 있다. 상기 컴퓨터 설비는 로봇, 프린터, 스캐너, 블랙박스, 네비게이터, 카메라, 비디오 카메라, 프로젝터, 손목 시계, 휴대용 저장기기, 웨어러블 설비, 교통수단, 가전 제품 및/또는 의료 기기에 적용될 수 있다. 여기서 교통수단은 비행기, 선박 및/또는 차량을 포함할 수 있고; 가전 제품은 텔레비전, 에어컨, 전자 레인지, 냉장고, 전기 밥솥, 가습기, 세탁기, 조명, 가스 렌지, 레인지 후드를 포함할 수 있고; 의료 기기는 자기공명 기기, 초음파 기기 및/또는 심전도 기기 등을 포함할 수 있다. 선택적으로, 상기 제 1 메모리(200)는 내장 메모리일 수 있고, 상기 제 2 메모리(300)는 하드 디스크 등과 같은 외장 메모리일 수 있다. 예를 들어, 상기 제 1 메모리(200)는 RAM(Random-Access Memory, 랜덤 액세스 제 1 메모리(200)) 들일 수 있고, 제 2 메모리(300)는 DDR(Double Data Rate, 더블 데이터 레이트 동기식 다이내믹 랜덤 제 1 메모리(200)) 등일 수 있다. 선택적으로, 상기 제 1 메모리(200)는 상기 프로세서(100)와 일체로 통합될 수 있고 , 즉 상기 제 1 메모리(200)는 캐시 메모리(Cache)와 같은 온-칩 메모리일 수 있으며, 상기 제 2 메모리(300)는 내장 메모리와 같은 오프-칩 메모리 일 수 있고, 예를 들어 RAM 등이다. 선택적으로, 제 2 메모리(300)는 컴퓨터 설비가 특정 연산을 수행하기 필요한 데이터 및 컴퓨터 프로그램 등을 저장하도록 구성된다. 또한, 상기 데이터는 신경망 데이터 등과 같은 기계 학습 데이터 일 수 있다. 제 1 메모리(200)의 저장 용

량이 보다 작기 때문에, 프로세서(100)가 특정 연산을 수행할 경우, 제 2 메모리(300)에 저장된 상기 특정 연산을 완료되기 위해 필요한 데이터를 제 1 메모리(200)에 기록하여, 프로세서(100)가 제 1 메모리(200)로부터 상기 특정 연산을 수행할 필요한 입력 데이터를 판독하여 연산을 수행할 수 있으며, 연산 결과를 제 1 메모리(200)에 기록한다. 일 실시예에서, 도 2를 참조하면, 해당 프로세서(100)는 제어기 유닛(110) 및 연산 유닛(110)을 포함할 수 있고, 여기서 제어기 유닛(110)과 연산 유닛(110)은 연결되고, 상기 연산 유닛(110)은 하나의 메인 프로세싱 회로(121) 및 복수의 슬레이브 프로세싱 회로(122)를 포함할 수 있고, 상기 메인 프로세싱 회로(121)와 슬레이브 프로세싱 회로(122)는 마스터-슬레이브 구조를 형성한다. 이에 대응하여, 상기 제 1 메모리(200)의 개수는 복수일 수 있고, 복수의 제 1 메모리(200)는 마스터-슬레이브 구조의 저장 시스템을 형성할 수 있다. 예를 들어, 복수의 제 1 메모리(200)는 하나의 메인 메모리와 복수의 슬레이브 메모리를 포함할 수 있고, 여기서, 상기 메인 메모리는 메인 프로세싱 회로에 인접하여 설치될 수 있고, 상기 슬레이브 메모리는 슬레이브 프로세싱 회로에 인접하여 설치될 수 있다. 선택적으로, 상기 메인 메모리는 메인 프로세싱 회로의 온-칩 메모리일 수 있으며, 상기 슬레이브 메모리는 슬레이브 프로세싱 회로의 온-칩 메모리 일 수 있다. 또한, 상기 메인 메모리의 저장 용량은 각 슬레이브 메모리의 저장 용량보다 작다. 또한, 슬레이브 프로세서 각각은 하나 이상의 슬레이브 메모리를 대응하여 설치할 수 있으며, 이에 구체적으로 제한되는 것은 아니다. 선택적으로, 전술한 제어기 유닛(110)은 데이터 및 계산 명령을 획득하도록 구성된다. 상기 데이터는 구체적으로 기계 학습 데이터를 포함할 수 있고, 선택적으로 상기 기계 학습 데이터는 신경망 데이터일 수 있다. 제어기 유닛(110)은 또한, 획득한 계산 명령을 분석하여 연산 명령을 획득하고, 복수의 연산 명령과 데이터를 메인 프로세싱 회로에 전송하도록 구성된다. 메인 프로세싱 회로(121)는 데이터 및 상기 메인 프로세싱 회로(121)와 복수의 슬레이브 프로세싱 회로(122) 사이에서 송신된 데이터 및 연산 명령에 대한 전처리를 수행하도록 구성된다. 복수의 슬레이브 프로세싱 회로(122)는 메인 프로세싱 회로(121)로부터 전송된 데이터 및 연산 명령에 따라 중간 연산을 수행하여 복수의 중간 결과를 획득하고, 복수의 중간 결과를 메인 프로세싱 회로(121)에 전송하도록 구성되며; 메인 프로세싱 회로(121)는 복수의 중간 결과에 대해 후속처리를 수행하여 계산 명령의 계산 결과를 획득하도록 구성된다. 상기 메인 프로세싱 회로(121)와 각 슬레이브 프로세싱 회로(122) 상에 모두 제 1 메모리가 집적되어 있고, 즉 복수의 제 1 메모리는 상기 메인 프로세싱 회로과 슬레이브 프로세싱 회로의 온-칩 메모리일 수 있고, 제 2 메모리는 상기 프로세서의 오프-칩 메모리일 수 있다.

[0083] 선택적으로, 상기 제어기 유닛(110)은 명령 캐시 유닛(111), 명령 처리 유닛(112) 및 저장 큐 유닛(114)을 포함할 수 있으며; 명령 캐시 유닛(111)은 기계 학습 데이터와 연관된 계산 명령을 저장하도록 구성되며; 명령 처리 유닛(112)은 계산 명령에 대한 분석을 통해 복수의 연산 명령을 획득하도록 구성되며; 저장 큐 유닛(114)은 명령 큐를 저장하도록 구성되며, 상기 명령 큐는 큐의 앞뒤 순서로 실행될 복수의 연산 명령 또는 계산 명령을 포함한다. 선택적으로, 상기 제어기 유닛(110)은 복수의 연산 명령들이 존재할 때, 제 1 연산 명령이 제 1 연산 명령 앞의 제 0 연산 명령과 연관되는지 여부를 확정하기 위한 의존관계 처리 유닛(113)을 포함할 수 있고, 제 1 연산 명령이 제 0 연산 명령과 연관 관계가 있을 경우, 제 1 연산 명령은 명령 저장 유닛에 캐싱하고, 제 0 연산 명령의 실행이 완료된 후 명령 저장 유닛으로부터 제 1 연산 명령을 추출하여 연산 유닛에 전송한다. 구체적으로, 의존관계 처리 유닛(113)이 제 1 연산 명령에 의해 제 1 연산 명령에서 필요한 데이터(예를 들어, 매트릭스)의 제 1 저장 주소 구간을 추출하고, 제0연산 명령에 의해 제0연산 명령에서 필요한 데이터의 제0 저장 주소 구간을 추출하다면, 제 1 저장 주소 구간과 제0저장 주소 구간 사이 겹치는 영역이 있을 경우, 제 1 연산 명령과 제0연산 명령은 연관 관계가 있다고 확정하고, 제 1 저장 주소 구간과 제0저장 주소 구간 사이 겹치는 영역이 없을 경우, 제 1 연산 명령과 제0연산 명령은 연관 관계가 없다고 확정한다.

[0084] 일 실시예에서, 도3에 제시된 바와 같이, 연산 유닛(120)은 브랜치 프로세싱 회로(123)을 더 포함할 수 있고, 여기서 메인 프로세싱 회로(121)와 브랜치 프로세싱 회로(123)는 연결되며, 브랜치 프로세싱 회로(123)는 복수의 슬레이브 프로세싱 회로(122)와 연결되고; 브랜치 프로세싱 회로(123)는 메인 프로세싱 회로(121)와 슬레이브 프로세싱 회로(122) 사이의 데이터 또는 명령을 전달하는 것을 수행하도록 구성된다. 본 실시예에서, 메인 프로세싱 회로(121)는 구체적으로 하나의 입력 신경원을 복수의 데이터 블록으로 분배하고, 복수의 데이터 블록 중의 적어도 하나의 데이터 블록, 가중치 및 복수의 연산 명령으로 이루어진 군으로분터 선택된 하나 이상의 연산 명령을 브랜치 프로세싱 회로에 전송하도록 구성되며; 브랜치 프로세싱 회로(123)는 메인 프로세싱 회로(121)와 복수의 슬레이브 프로세싱 회로(122) 사이의 데이터 블록, 가중치 및 연산 명령을 전달하는 데 사용되고; 복수의 슬레이브 프로세싱 회로(122)는 해당 연산 명령에 따라 획득된 데이터 블록 및 가중치에 대해 연산을 실행하여 중간 결과를 얻고, 그 중간 결과를 브랜치 프로세싱 회로(123)에 전송하도록 구성되며; 메인 프로세싱 회로(121)은 브랜치 프로세싱 회로에 의해 전송된 중간 결과에 대한 후속 처리를 진행하여 해당 계산 명령의 결과를 얻고, 해당 계산 명령의 결과는 상기 제어기 유닛으로 전송하도록 구성된다. 선택적으로, 각 브랜치

프로세싱 회로(123) 상에도 제 1 메모리가 집적되어 있다.

[0085] 또 다른 실시예에서, 도 4에 제시된 바와 같이, 연산 유닛(120)은 하나의 메인 프로세싱 회로(121)와 복수의 슬레이브 프로세싱 회로(122)를 포함할 수 있다. 여기서, 복수의 슬레이브 프로세싱 회로는 어레이 형태로 분포되어 있고; 각 슬레이브 프로세싱 회로는 인접한 기타 슬레이브 프로세싱 회로와 연결되고, 메인 프로세싱 회로는 복수의 슬레이브 프로세싱 회로 중의 k개의 슬레이브 프로세싱 회로와 연결되고, k개의 슬레이브 프로세싱 회로는: 제 1 행의 n 개의 슬레이브 프로세싱 회로, 제 m 행의 n 개 슬레이브 프로세싱 회로 및 제 1 열의 m 개 슬레이브 프로세싱 회로이다. 도 1c에 표시된 바와 같이, k개의 슬레이브 프로세싱 회로는 제 1 행의 n 개 슬레이브 프로세싱 회로, 제 m 행의 n 개 슬레이브 프로세싱 회로 및 제 1열의 m 개 슬레이브 프로세싱 회로만 포함하고, 즉 상기 k개의 슬레이브 프로세싱 회로는 복수의 슬레이브 프로세싱 회로 중의 메인 프로세싱 회로에 직접적으로 연결된 슬레이브 프로세싱 회로이다. K개의 슬레이브 프로세싱 회로는 메인 프로세싱 회로와 복수의 슬레이브 프로세싱 회로 사이의 데이터 및 명령을 전달하도록 구성된다.

[0086] 본 개시에 의해 제공되는 프로세서는 연산 유닛를 원 마스터-멀티 슬레이브 구조로 설치되여, 포워드 연산의 계산 명령에 대해, 포워드 연산의 계산 명령에 따라 데이터를 분할할 수 있고, 복수의 슬레이브 프로세싱 회로가 계산량이 많은 부분에 대해 병렬 연산을 진행함으로써, 연산 속도가 향상되고 연산 시간이 절약되어 전력 소모를 감소한다.

[0087] 선택적으로, 상기 기계 학습 계산은 인공 신경망 연산을 포함할 수 있으며, 상기 입력 데이터는 입력 신경원 데이터 및 가중치 데이터를 포함할 수 있다. 상기 계산 결과는 구체적으로 인공 신경망 연산의 결과일 수 있고, 즉 출력 신경원 데이터일 수 있다.

[0088] 신경망에서, 연산은 신경망 중 한층의 연산일 수 있고, 다층 신경망의 실현 과정으로서 포워드 연산인 경우, 윗층의 인공 신경망 수행이 완료된 후, 다음 층의 연산 명령은 연산 유닛 내 산출된 출력 신경원을 다음 층의 입력 신경원으로 사용하여 연산을 진행하고(또는 상기 출력 신경원에 대해 특정 동작을 진행 후 다음 층의 입력 신경원으로 사용함), 한편 가중치도 다음 층의 가중치로 대체되며; 백워드 연산인 경우, 윗 층의 인공 신경망의 백워드 연산이 완료된 후, 다음 층의 연산 명령은 연산 유닛 내 산출된 입력 신경원 그래디언트를 다음 층의 출력 신경원 그래디언트로 사용하여 연산을 진행하고(또는 상기 입력 신경원 그래디언트에 대해 특정 동작을 진행 후 다음 층의 출력 신경원으로 사용함), 한편 가중치 또한 다음 층의 가중치로 대체한다.

[0089] 상기 기계 학습 계산은 추가로 서포트 벡터 머신 연산, K-근접(K-nn) 연산, K-평균치(K-means) 연산, 주성분 분석 연산 등을 포함할 수 있다. 설명의 편의를 위해 다음은 인공 신경망 연산을 예로 기계 학습 계산의 상세 방안을 설명한다.

[0090] 인공 신경망 연산에 대해, 상기 인공 신경망 연산이 다층의 연산을 갖고 있을 경우, 다층 연산의 입력 신경원과 출력 신경원은 전체 인공 신경망의 입력층 신경원과 출력층 신경원이 아니라, 네크워크 내의 임의 인접한 두 층에서 네트워크 포워드 연산의 하위 층에 있는 신경원은 입력 신경원이고, 네트워크 포워드 연산의 상위 층에 있는 신경원은 출력 신경원이다. 컨벌루션 신경망을 예로 들어, 하나의 컨벌루션 신경망이 L층이 있을 경우, K=1,2,…,L-1, 제K층과 제K+1층에 있어서, 제K층을 입력층으로 지칭하고, 이 중의 신경원은 상기 입력 신경원이며, 제K+1층을 출력층으로 지칭하고, 이 중의 신경원은 상기 출력 신경원이다. 즉, 최상위 층을 제외하고 모든 층은 입력층로 사용할 수 있으며, 그 다음 층은 대응된 출력층이다.일 실시예에서, 제 2 메모리는 컴퓨터 프로그램을 저장하도록 구성되며, 상기 프로세서가 상기 컴퓨터 프로그램을 실행할 경우, 본 개시 실시예의 데이터 전처리 방법을 구현할 수 있어, 상기 처리하고자 하는 연산의 수행 과정에서 각 데이터의 저장 공간 분배 규칙을 얻을 수 있다. 구체적으로, 상기 컴퓨터 설비는 다음과 같은 데이터 전처리 방법을 수행하도록 구성될 수 있으며, 처리하고자 하는 연산(예를 들면, 신경망 연산 등)에 대해 전처리를 수행하여, 제 1 메모리 상에서 상기 처리하고자 하는 연산의 입력 데이터, 출력 데이터 및 중간 계산 결과 등 데이터의 데이터 저장 공간 분배 규칙을 획득한다. 따라서, 상기 프로세서가 해당 처리하고자 하는 연산을 실행할 때, 상기 처리하고자 하는 연산에 관련된 데이터(입력 데이터, 출력 데이터, 중간 결과 등)는 상기 저장 공간 분배 규칙에 따라 제 1 메모리에 저장할 수 있다. 이와 같이, 연산 과정에서 저장 자원을 미리 분배함으로써, 제 1 메모리의 저장 공간을 합리적으로 이용할 수 있을 뿐만 아니라, 처리의 연산 속도 및 정밀도를 향상시킬 수 있다. 여기서, 상기 저장 공간 분배 규칙은 처리하고자 하는 연산을 수행하는 과정에서의 입력 데이터의 저장 주소, 출력 데이터의 저장 주소, 중간 계산 결과의 저장 주소 및 각 저장 공간내 저장된 데이터의 업데이트 규칙 등을 포함할 수 있다. 구체적인 내용은 아래 설명을 참조할 수 있다. 본 개시의 실시예에서는, 연산 과정에서의 데이터 리드-라이트 동작(즉, I/O 동작 횟수의 감소)을 감소시키기 위한, 데이터 전처리 방법을 제공하며, 상기 데이터 전처리 방법은 상기

컴퓨터 설비에 적용될 수 있다. 구체적으로, 도 5에 제시된 바와 같이, 상기 데이터 전처리 방법은 하기 단계를 포함할 수 있다 :

[0091] S100단계: 제 1 메모리의 가용 저장 용량 및 목표 연산 동작을 얻는다.

[0092] 구체적으로, 프로세서는 상기 제 1 메모리의 구성 정보(예를 들어, 상기 제 1 메모리의 모델번호 등 정보)에 따라 상기 제 1 메모리의 총 저장 용량을 획득할 수 있다. 또한, 프로세서는 상기 제 1 메모리의 총 저장 용량 및 상기 제 1 메모리상의 점유된 저장 용량에 따라, 상기 제 1 메모리의 가용 저장 용량을 획득할 수 있다.

[0093] 본 개시의 실시예에서, 프로세서는 처리하고자 하는 연산을 획득하고, 상기 처리하고자 하는 연산 및 제 1 메모리의 가용 저장 용량에 따라 목표 연산 동작을 확정할 수 있다. 여기서, 상기 처리하고자 하는 연산은 하나 이상의 연산 동작을 포함할 수 있고, 상기 처리하고자 하는 연산은 신경망과 같은 연산일 수 있다. 예를 들어, 상기 처리하고자 하는 연산은 덧셈 동작, 뺄셈 동작, 곱셈 동작, 나눗셈 동작, 컨벌루션 동작, 풀링(Pooling) 동작 및 활성화(예 : Relu) 동작 등의 연산 동작을 포함할 수 있으며, 여기서 구체적인 제한은 하지 않는다. 상기 목표 연산 동작은 처리하고자 하는 연산에서에서 하나 이상의 연산 동작의 조합 일 수 있다.

[0094] S200단계: 목표 연산 동작 및 제 1 메모리의 가용 저장 용량에 따라, 목표 연산 동작에 대응하는 목표 입력 데이터를 확정한다. 여기서, 목표 입력 데이터는 목표 연산 동작에 대응하는 모든 입력 데이터의 일부 또는 전부이다.

[0095] 구체적으로, 프로세서는 상기 목표 연산 동작에 따라 상기 목표 연산 동작이 완료되기 필요한 모든 입력 데이터 및 상기 전체 입력 데이터의 데이터 용량(즉, 상기 전체 입력 데이터에 수용되는 저장 공간 크기)을 확정할 수 있다. 또한, 프로세서는 제 1 메모리의 가용 저장 용량 및 상기 목표 연산 동작의 모든 입력 데이터의 데이터 용량에 따라, 상기 목표 연산 동작에 대응하는 목표 입력 데이터 및 그 데이터 용량을 확정할 수 있고, 상기 목표 입력 데이터의 데이터 용량은 제 1 메모리의 저장 용량보다 작거나 같다. 여기서, 상기 목표 입력 데이터는 목표 연산 동작에 대응하는 모든 입력 데이터의 일부 또는 전부이고, 즉 상기 목표 입력 데이터의 데이터 용량은 목표 연산 동작에 대응하는 모든 입력 데이터의 데이터 용량보다 작거나 같다. 목표 입력 데이터의 데이터 용량이 목표 연산 동작에 대응하는 모든 입력 데이터의 데이터 용량보다 작을 경우, 상기 목표 연산 동작에 대응하는 모든 입력 데이터의 일부만 제 1 메모리 상에 로딩함으로써, 제 1 메모리 상에 일정 일부의 저장 공간이 사전 보류되어, 상기 목표 연산 동작의 목표 출력 데이터 및 중간 계산 결과 등과 같은 데이터를 저장하도록 구성될 수 있다. 목표 입력 데이터의 데이터 용량이 목표 연산 동작에 대응하는 모든 입력 데이터의 데이터 용량과 같을 경우, 저장 공간에 대한 중복 사용을 통해 상기 목표 연산 동작의 목표 출력 데이터 및 중간 계산 결과 등과 같은 데이터를 저장할 수 있다.

[0096] S300단계: 목표 연산 동작 및 목표 입력 데이터에 따라, 목표 연산 동작의 목표 출력 데이터를 확정한다.

[0097] 구체적으로, 처리하고자 하는 연산의 계산량은 정적 분석 가능하기에, 프로세서는 상기 목표 연산 동작의 목표 입력 데이터 및 상기 목표 연산 동작에 따라, 상기 목표 연산 동작의 목표 출력 데이터 및 상기 목표 출력 데이터의 데이터 용량 등 정보를 얻을 수 있고, 즉 프로세서는 상기 목표 연산 동작의 목표 출력 데이터에 필요한 저장 공간을 얻을 수 있다.

[0098] S400단계: 목표 연산 동작의 목표 출력 데이터가 상기 목표 연산 동작 후의 기타 연산 동작의 입력 데이터인 경우, 상기 목표 출력 데이터가 제 1 메모리 상에 저장되어, 목표 출력 데이터의 판독 횟수를 줄인다.

[0099] 구체적으로, 상기 목표 연산 동작의 목표 출력 데이터가 상기 목표 연산 동작 후의 기타 연산 동작의 입력 데이터인 경우, 즉 상기 목표 연산 동작 후 여전히 상기 목표 출력 데이터가 필요할 경우, 상기 목표 출력 데이터를 상기 제 1 메모리 상에 저장할 수 있고, 목표 출력 데이터의 판독 횟수를 감소시킴으로써 프로세서의 속도 및 효율을 향상시킨다.

[0100] 종래의 기술에서, 프로세서가 상기 목표 연산 동작을 수행하여 상기 목표 출력 데이터를 얻은 후, 프로세서는 상기 목표 출력 데이터를 제 1 메모리에서 제 2 메모리로 전송함으로써, 상기 목표 출력 데이터가 제 1 메모리에서 점유했던 저장 공간을 해제한다. 상기 목표 연산 동작 후의 연산 동작은 상기 목표 출력 데이터가 계속 사용할 필요가 있을 경우, 프로세서는 상기 목표 출력 데이터를 제 2 메모리에서 제 1 메모리에 다시 전송할 필요가 있고, 이러한 방법은 목표 출력 데이터의 I/O 판독 동작을 여러번 수행할 필용가 있음으로써, 연산 시간의 증가를 초래하기 쉬워서, 프로세서의 효율 및 속도가 보다 낮게 된다. 본 개시 실시예의 데이터 전처리 방법은, 종래의 기술에 비하여 목표 출력 데이터의 판독 횟수를 감소시킴으로써, 연산 과정에서의 I/O 판독 동작의 점유 시간을 단축할 수 있어, 프로세서의 속도 및 효율을 향상시킬 수 있다. 예를 들어, 도 8에 제시된 바와 같이,

프로세서는 목표 연산 동작 OP1을 얻을 수 있고, 상기 목표 연산 동작 OP1의 모든 입력 데이터는 입력 데이터 X이다(X는 서브 입력 데이터 X11, X21, X12 및 X22를 포함하며, 여기서, 서브 입력 데이터X11 및 X12는 입력 데이터X1을 구성할 수 있고, 서브 입력 데이터X21 및 X22는 입력 데이터X2를 구성할 수 있고, 상기 입력 데이터 X1 및 X2는 벡터 또는 매트릭스 데이터 등일 수 있다). 프로세서는 상기 목표 연산 동작 OP1과 제 1 메모리의 가용 저장 용량에 따라, 서브 입력 데이터 X11 및 X21을 상기 목표 연산 동작 OP1의 목표 입력 데이터로 사용할 수 있다. 또한, 상기 프로세서는 목표 연산 동작(OP1) 및 목표 입력 데이터 X11 및 X21에 따라 목표 출력 데이터 Y1 및 상기 목표 출력 데이터 Y1의 데이터 용량을 확정할 수 있다. 또한, 프로세서는 설정된 연산 규칙에 따라, 목표 연산 동작 OP1 후의 기타 연산 동작에서 상기 목표 출력 데이터 Y1을 사용할 필요한지 여부를 확정할 수 있고, 목표 연산 동작 OP1 후의 기타 연산 동작에서 상기 목표 출력 데이터 Y1이 사용될 필요가 있을 경우, 상기 목표 출력 데이터(Y1)이 목표 연산 동작 OP1 후의 기타 연산 동작(OP2)의 입력 데이터이다면, 상기 목표 출력 데이터Y1은 제 1 메모리 상에 일시적으로 저장된다. 따라서, 다음 목표 연산 동작이 연산 동작 OP2일 경우, 프로세서는 다음 연산 동작 OP2를 수행하기 전에, 설정된 규칙에 따라 상기 연산 동작 OP2에 필요한 입력 데이터 Y3만 제 2 메모리로부터 제 1 메모리에 전송하면 되고, 더 이상 상기 목표 출력 데이터 Y1의 전송 단계를 수행할 필요가 없다. 또한, 상기 목표 출력 데이터 Y1은 목표 연산 동작 OP1 후의 연산 동작 OP2의 입력 데이터인 경우, 한편 상기 목표 출력 데이터 Y1은 연산 동작 OP3의 입력 데이터이다. 이런 경우, 상기 목표 출력 데이터 Y1을 제 1 메모리 상에 저장하고, 연산 동작 OP2와 OP3이 완료된 다음, 상기 목표 출력 데이터 Y1을 제 1 메모리에서 삭제하여 목표 출력 데이터 Y1가 점유하였던 제 1 메모리의 저장 공간을 해제할 수 있다.본 개시 실시예의 데이터 전처리 방법은, 연산 동작OP1의 계산 종료후, 목표 출력 데이터 Y1을 제 1 메모리로부터 제 2 메모리에 전송하는 과정. 및 연산 동작 OP2를 수행할 때 목표 출력 데이터 Y1을 다시 제 2 메모리로부터 제 1 메모리에 전송하는 과정을 감소시켜, 목표 출력 데이터의 판독 횟수를 감소시킴으로써, 연산 과정에서의 I/O 판독 동작의 점유 시간을 단축할 수 있어, 프로세서의 속도 및 효율을 향상시킬 수 있다. 선택적으로, 전술한 처리하고자 하는 연산은 복수의 연산 계층을 포함하는 신경망 연산일 수 있으며, 도8에 제시된 바와 같이, 상기 연산 동작 OP1 및 OP2는 신경망 연산 중의 연산 계층일 수 있다. 전술힌 입력 데이터 X는 입력 신경원 데이터 및 가중치 데이터 등을 포함할 수 있고, 그는 입력 데이터 X1 및 X2를 포함할 수 있다. 선택적으로, 상기 입력 데이터 X1 및 X2는 각각 기타 연산 계층에 속할 수 있다. 또한, 프로세서는 상기 목표 연산 계층(OP1) 및 제 1 메모리의 가용 저장 용량에 따라, 서브 입력 데이터X11 및 X21)를 상기 목표 연산 계층(OP1)의 목표 입력 데이터로 사용할 수 있다. 또한, 상기 프로세서는 목표 연산 계층(OP1) 및 목표 입력 데이터 X11과 X21에 따라, 목표 출력 데이터 Y1 및 목표 출력 데이터 Y1의 데이터 용량을 확정할 수 있고, 상기 목표 출력 데이터 Y1은 연산 계층(OP1)의 출력 데이터의 일부이며, 상기 출력 데이터는 연산 계층(OP1)의 출력 신경원 데이터 및 가중치 등을 포함할 수 있다.기타 예로, 도 9에 제시된 바와 같이, 상기 처리하고자 하는 연산은 신경망 연산이고, 상기 처리하고자 하는 연산은 컨벌루션 계층, 풀링 계층 및 활성화 계층을 포함할 수 있으며, 전술한 각 연산 계층의 실행 순서는 차례로 컨벌루션 연산 동작 - 풀링 연산 동작 - 활성화 연산 동작이다. 즉, 컨벌루션 연산 동작의 출력 데이터는 풀링 연산 동작의 입력 데이터이고, 풀링 연산 동작의 출력 데이터는 활성화 연산 동작의 입력 데이터이다. 각 연산 계층의 입력 데이터는 상기 연산 계층에 대응하는 입력 신경원 데이터 및 가중치 등 데이터를 포함할 수 있다. 현재 목표 연산 동작이 풀링 연산 동작인 경우, 프로세서는 제 1 메모리의 가용 저장 용량 및 목표 연산 동작에 따라, 상기 풀링 연산 동작에 대응하는 C1-C2구간내의 데이터인 목표 입력 데이터를 얻을 수 있다(C1-C2구간내 데이터는 컨벌루션 연산 동작의 출력 데이터를 나타내며, 컨벌루션 연산 동작에 대응하는 출력 신경원 데이터 및 가중치 등을 포함할 수 있다). 상기 목표 입력 데이터 C1-C2에 대응하는 목표 출력 데이터는 B1-B2구간내의 데이터이다(여기서, B1-B2구간내의 목표 출력 데이터는 풀링 연산 동작에 대응하는 출력 신경원 데이터 및 가중치 등을 포함할 수 있다). 또한, 상기 풀링 연산 동작에 대응하는 목표 출력 데이터(B1-B2)가 활성화 연산 동작의 입력 데이터이기 때문에, 상기 풀링 연산 동작의 목표 출력 데이터(B1-B2)를 제 1 메모리에 저장할 수 있다. 따라서, 풀링 연산 동작이 완료된 후에, 목표 출력 데이터(B1-B2)를 제 1 메모리로부터 제 2 메모리로 운반하여 제 1 메모리상의 저장 공간을 해제할 필요가 없다. 또한, 활성화 연산 동작을 수행하기 전에 목표 출력 데이터(B1-B2)를 제 2 메모리로부터 제 1 메모리로 다시 전송할 필요도 없다.종래 기술에서, 프로세서가 목표 출력 데이터(B1-B2)를 획득한 후에, 우선 상기 목표 출력 데이터(B1-B2)를 제 1 메모리로부터 제 2 메모리로 전송하여 제 1 메모리의 저장 공간을 해제시킨다. 활성화 연산 동작의 입력 데이터는 풀링 연산 동작의 출력 데이터에 의존하기 때문에, 프로세서가 활성화 연산 동작을 수행하기 전 상기 풀링 연산 동작에 대응하는 목표 출력 데이터B1-B2인 데이터 블록을 다시 제 2 메모리로부터 제 1 메모리로 전송할 수 있다. I/O 대역폭이 제한된 경우, 이와 같은 빈번한 데이터 판독 동작은 프로세서의 처리 효율성에 영향을 미친다. 따라서, 현존 기술에 비해, 본 개시 실시예의 데이터 전처리 방법으로서, 목표 출력 데이터의 판독 횟수를 감소시킴으로써(즉 목표 출력 데이터의 load 및 store 동작을 감소시킨다), 연산 과정에서의 I/O 판독 동

작의 점유 시간을 단축할 수 있어, 프로세서의 속도 및 효율을 향상시킬 수 있다.

[0101] 일 실시예에서, 상기 방법은 추가로 하기 단계를 포함한다:

[0102] 목표 연산 동작의 목표 출력 데이터가 목표 연산 동작 후의 기타 연산 동작의 입력 데이터인 경우(즉, 상기 목표 연산 동작의 목표 출력 데이터가 상기 처리하고자 하는 연산의 중간 결과 데이터이다), 상기 목표 연산 동작의 목표 출력 데이터를 제 1 메모리 또는 제 1 메모리와 제 2 메모리 상에 저장한다. 구체적으로, 목표 연산 동작의 목표 출력 데이터가 목표 연산 동작 후의 기타 연산 동작의 입력 데이터인 경우, 목표 출력 데이터를 제 1 메모리에 저장하여, 상기 목표 출력 데이터의 중복 로딩 동작을 감소시킬 수 있다(즉, 목표 출력 데이터의 load 동작을 감소). 한편, 상기 목표 출력 데이터를 제 1 메모리로부터 제 2 메모리로 복사할 수 있어서, 제 1 메모리와 제 2 메모리 상의 데이터 일관성을 보장한다. 선택적으로, 상기 목표 연산 동작에 대응하는 목표 출력 데이터를 제 2 메모리에 동기화 저장할 필요가 있는지 여부는 구체적인 연산 수요에 따라 결정될 수 있다.상기 목표 출력 데이터를 제 2 메모리에 동기화 저장하는 것이 불필요한 경우, 상기 목표 출력 데이터를 제 1 메모리에만 저장함으로써, 목표 출력 데이터의 load 및 store 동작을 함께 감소시킬 수 있다. 상기 목표 출력 데이터를 제 2 메모리에 동기화 저장이 필요할 경우, 목표 출력 데이터를 제 1 메모리와 제 2 메모리 상에 동기화 저장할 수 있고, 상기 목표 출력 데이터의 load 동작을 감소킴으로써, 데이터의 판독 동작이I/O 대역폭을 과도하게 점유하여 프로세서의 처리 속도에 영향을 미치는 경우를 피할 수 있다.도8을 참조하면, 목표 연산 동작 OP1 후의 기타 연산 동작에서 상기 목표 출력 데이터 Y1를 사용할 필요가 있을 경우, 상기 목표 출력 데이터(Y1)이 목표 연산 동작 OP1 후의 기타 연산 동작(OP2)의 입력 데이터라면, 상기 목표 출력 데이터(Y1)는 제 1 메모리 상기에 일시적으로 저장된다. 따라서, 다음 목표 연산 동작이 연산 동작 OP2일 경우, 프로세서는 다음 연산 동작 OP2를 수행하기 전에, 설정된 규칙에 따라 상기 연산 동작 OP2에서 필요한 입력 데이터 Y3만 제 2 메모리로부터 제 1 메모리에 전송하고, 더 이상 상기 목표 출력 데이터(Y1)에 대한 전송 단계를 수행할 필요가 없다. 또한, 프로세서는 제 1 메모리 및 제 2 메모리의 데이터가 일치하도록 목표 출력 데이터(Y1)를 제 1 메모리로부터 제 2 메모리에 복사할 수도 있다. 따라서, 본 개시 실시예의 데이터 전처리 방법은, 연산 동작OP1의 계산 종료 후에, 목표 출력 데이터 Y1을 제 1 메모리로부터 제 2 메모리에 전송하는 과정을 감소시켜, 목표 출력 데이터의 판독 횟수를 감소시킴으로써, 연산 과정에서의 I/O 판독 동작의 점유 시간을 단축할 수 있어, 프로세서의 속도 및 효율을 향상시킬 수 있다.도9를 참조하면, 상기 풀링 연산 동작에 대응하는 목표 출력 데이터 B1-B2가 활성화 연산 동작의 입력 데이터이기 때문에, 상기 풀링 연산 동작에 대응하는 목표 출력 데이터 B1-B2를 함께 제 1 메모리와 제 2 메모리 상에 저장할 수 있다. 따라서, 활성화 연산 동작을 수행하기 전에 상기 목표 출력 데이터 B1-B2를 제 2 메모리로부터 제 1 메모리로 다시 전송할 필요가 없다. 한편, 풀링 연산 동작이 완료된 후에, 목표 출력 데이터B1-B2가 제 1 메모리로부터 제 2 메모리로 복사되어 제 1 메모리와 제 2 메모리의 데이터의 일관성을 보장할 수 있다. 현존 기술에 비해, 본 개시 실시예의 데이터 전처리 방법은, 목표 출력 데이터 B1-B2를 제 2 메모리로부터 제 1 메모리에 재전송하는 과정을 감소시켜, 목표 출력 데이터의 판독 횟수를 감소시킴으로써, 연산 과정에서의 I/O 판독 동작의 점유 시간을 단축할 수 있어, 프로세서의 속도 및 효율을 향상시킬 수 있다. 일 실시예에서, 처리하고자 하는 연산의 각 목표 연산 동작에 필요한 모든 입력 데이터의 데이터 용량이 모두 보다 크기 때문에, 프로세서는 각 목표 연산 동작과 관련된 모든 입력 데이터를 분할할 수 있다. 즉, 제 1 메모리의 가용 저장 용량에 따라, 각 목표 연산 동작에 관련된 모든 입력 데이터(입력 신경원 데이터 및 가중치 등을 포함함)를 복수의 입력 데이터 블록으로 분할할 수 있고, 각 입력 데이터 블록에 대해 상기 목표 연산 동작을 각각 수행하여, 해당 목표 연산 동작의 계산 결과를 얻는다. 마지막으로, 각 입력 데이터 블록에 대응하는 계산 결과를 융합시킴으로써, 상기 목표 연산 동작에 대응하는 출력 데이터를 얻을 수 있다. 상기 입력 데이터 블록은 바로 상기 목표 입력 데이터이고, 각 입력 데이터 블록에 대응하는 출력 데이터는 바로 상기 목표 출력 데이터이다. 선택적으로, 전술한 S200 단계에 하기 내용이 구체적으로 포함된다: 프로세서는 제 1 메모리의 가용 저장 용량과 상기 목표 연산 동작에 필요한 전부 입력 데이터의 데이터 용량에 따라, 상기 목표 연산 동작에 대응하는 입력 데이터 블록을 확정할 수 있고, 상기 입력 데이터 블록을 상기 목표 연산 동작에 대응하는 목표 입력 데이터로 사용한다. 구체적으로는, 상기 목표 연산 동작에 필요한 전부 입력 데이터의 데이터 용량이 상기 제 1 메모리의 가용 저장 용량보다 큰 경우, 프로세서는 제 1 메모리의 가용 저장 용량에 따라 상기 목표 연산 동작에 대응하는 입력 데이터 블록을 확정할 수 있고, 상기 입력 데이터 블록은 상기 목표 연산 동작의 전부 입력 데이터의 일부이다. 상기 목표 연산 동작에 필요한 전부 입력 데이터의 데이터 용량이 상기 제 1 메모리의 가용 저장 용량보다 작거나 같은 경우, 상기 목표 연산 동작의 전부 입력 데이터를 하나의 입력 데이터 블록으로 사용할 수 있고, 즉 상기 목표 연산 동작의 전부 입력 데이터를 목표 입력 데이터로 사용한다.

[0103] 예를 들어, 도 8에 제시된 바와 같이, 프로세서는 현재의 목표 연산 동작 OP1을 얻을 수 있고, 상기 목표 연산 동작 OP1의 모든 입력 데이터는 모든 입력 데이터 X(X는 입력 데이터 X1및 X2를 포함)이다. 프로세서는, 상기

목표 연산 동작 OP1과 제 1 메모리의 가용 저장 용량에 따라, 상기 서브 입력 데이터 X21 및 상기 서브 입력 데이터 X21을 상기 목표 연산 동작 OP1의 목표 입력 데이터로 사용하고, 여기서 상기 서브 입력 데이터 X11과 서브 입력 데이터 X21의 데이터 용량의 합은 제 1 메모리의 가용 저장 용량보다 작다. 물론, 기타 실시예에서, 상기 목표 연산 동작에 대응하는 모든 입력 데이터 X의 데이터 용량이 제 1 메모리의 가용 저장 용량보다 작은 경우, 상기 목표 연산 동작에 대응하는 모든 입력 데이터를 모두 제 1 메모리에 로딩할 수 있다. 기타 예로서, 도 9에 제시된 바와 같이, 현재 목표 연산 동작이 풀링 연산 동작인 경우, 프로세서는 목표 연산 동작제 1 메모리의 가용 저장 용량 및 목표 연산 동작에 따라, C1-C2구간 내의 데이터(C1-C2구간 내의 데이터는 컨벌루션 연산 동작의 출력 데이터를 나타낸다) 를 하나의 입력 데이터 블록으로 사용할 수 있고, 상기 입력 데이터 블록을 상기 풀링 연산 동작에 대응하는 목표 입력 데이터로 사용한다. 현재 목표 연산 동작이 활성화 연산 동작 인 경우, 프로세서는 제 1 메모리의 가용 저장 용량에 따라, B1-B2구간 내의 데이터를 상기 활성화 연산 동작의 입력 데이터 블록으로 사용할 수 있고, 상기 입력 데이터 블록을 상기 활성화 연산 동작의 목표 입력 데이터로 사용한다. 일 실시예에서, 각 목표 연산 동작에 관련된 전부 입력 데이터가 복수의 입력 데이터 블록들로 분할되면, 각 입력 데이터 블록의 데이터 용량은 제 1 메모리의 저장 용량보다 작기 때문에, 상기 목표 연산 동작은 처리하고자 하는 연산의 복수의 연산 동작을 융합할 수 있어서, 제 1 메모리의 저장 공간을 최대한 이용하고 연산의 효율을 향상시킨다. 선택적으로, 상기 목표 연산 동작은 하나 이상의 연산 동작을 포함하고, 즉 상기 목표 연산 동작은 하나 이상의 연산 동작의 조합이다. 일반적으로, 상기 목표 연산 동작에 포함되는 각 연산 동작은 서로 기타 연산을 구현하기 위한 기타 연산 동작이다. 여기서, 프로세서는 제 1 메모리의 가용 저장 용량에 따라 각 연산 동작에 대응하는 서브 목표 입력 데이터를 확정할 수 있고, 각 연산 동작에 대응하는 서브 목표 입력 데이터에 따라 상기 목표 연산 동작에 대응하는 목표 입력 데이터를 확정할 수 있다. 구체적으로, 도 6을 참조하면, 상기 S200단계에서 목표 연산 동작에 대응하는 입력 데이터 블록을 확정하는 단계에 추가로 하기 단계를 포함한다.

[0104] S210단계: 제 1 메모리의 가용 저장 용량 및 각 연산 동작의 융합 속성에 따라, 융합 가능한 연산 동작의 수량을 확정하고 융합 수량 역치를 얻는다. 여기서, 각 연산 동작의 융합 속성은 각 연산 동작과 관련된 입력 데이터 및/또는 출력 데이터 간의 데이터 의존관계를 포함할 수 있다. 하나 또는 복수의 연산 동작이 프로세서에 의해 함께 수행될 수 있을 겨우, 상기 하나 또는 복수의 연산 동작은 융합될 수 있고, 융합도가 비교적 높다고 간주한다. 하나 또는 복수의 연산 동작이 프로세서에 의해 함께 수행될 수 없을 경우, 상기 하나 또는 복수의 연산 동작은 융합될 수 없고 융합도가 낮다고 간주한다. 각 연산 동작 간의 융합도는 미리 설정된 연산 규칙에 의해 결정될 수 있으며, 여기서 구체적인 제한은 하지 않는다.

[0105] S220단계: 하나 이상의 연산을 융합할 수 있고, 수량이 선정된 조합을 하나의 목표 연산 동작으로 사용하고, 여기서, 상기 선정된 수량은 융합 수량 역치 이하이다.

[0106] 예를 들어, 상기 선정된 수량이 융합 수량 역치와 같고 제 1 메모리의 저장 용량에 따라 확정한 융합될 수 있는 복수의 연산 동작은 하나의 목표 연산 동작인 것과 같다.

[0107] S230단계: 상기 수량이 선정된 연산 동작 각각에 대응하는 서브 목표 입력 데이터를 상기 목표 연산 동작에 대응하는 목표 입력 데이터로 사용한다.

[0108] 예를 들어, 도 8에 제시된 바와 같이, 처리하고자 하는 연산은 연산 동작 OP1 및 OP2를 포함할 수 있고, 2개의 연산 동작의 융합 속성에 따라, 상기 연산 동작 OP1 및 OP2는 프로세서에 의해 함께 실행될 수 있고, 제 1 메모리의 가용 저장 용량이 연산 동작 OP1의 목표 입력 데이터와 목표 출력 데이터, 및 연산 동작 OP2의 입력 데이터와 출력 데이터를 저장할 수 있을 경우, 상기 목표 연산 동작의 융합 가능한 연산 동작의 수량은 2개로 일 수 있으며, 여기서, 상기 연산 동작 OP1 및 OP2를 하나의 목표 연산 동작으로 사용할 수 있다. 한편, 상기 연산 동작 OP1 및 OP2에 대응하는 서브 목표 입력 데이터 X11, X21 및 Y3은 상기 목표 연산 동작의 목표 입력 데이터로 사용된다. 연산 동작 OP1 및 OP2가 융합될 수는 있지만, 제 1 메모리의 가용 저장 용량이 연산 동작 OP1의 목표 입력 데이터와 목표 출력 데이터만 수용할 수 있고, 연산 동작 OP2의 입력 데이터와 출력 데이터는 전부 수용할 수 없을 경우, 상기 목표 연산 동작의 융합 가능한 연산 동작의 수량은 1개로 일 수 있고, 여기서, 상기 연산 동작 OP1을 하나의 목표 연산 동작으로 사용할 수 있다. 한편, 상기 연산 동작 OP1에 대응하는 서브 목표 입력 데이터 X11, X21은 상기 목표 연산 동작의 목표 입력 데이터로 사용된다. 물론, 기타 실시예에서, 상기 목표 연산 동작은 2개 이상의 연산 동작을 포함할 수 있다. 예를 들면, 상기 처리하고자 하는 연산의 깊이 방향에서, 상기 연산 동작 OP2 후에 융합 가능한 기타 연산 동작이 존재하고, 상기 융합 가능한 연산 동작에 대응하는 목표 입력 데이터 및 목표 출력 데이터의 데이터 용량이 제 1 메모리의 가용 저장 용량에 만족할 수 있을 경우, 상기 목표 연산 동작에 포함된 연산 동작의 수량은 OP1, OP2 및 OPn(n은 2보다 크고, n은 자연수이다)일 수 있

다. 여기서, OP1, OP2 및 OPn에 대응하는 목표 입력 데이터의 데이터 및 목표 출력 데이터의 데이터 용량의 합은 제 1 메모리의 가용 저장 용량보다 작거나 같다. 또한, 상기 처리하고자 하는 연산은 복수의 연산 계층을 포함하는 신경망 연산일 수 있으며, 각 연산 계층은 하나의 연산 동작을 나타낼 수 있다. 예를 들면, 프로세서가 신경망 등에 대해 연산을 진행하는 경우, 신경망의 각 연산 계층은 모두 하나의 연산 동작일 수 있고, 상기 신경망의 각 연산 계층의 연결 관계에 따라 각 연산 동작의 융합 속성을 확정할 수 있고, 즉 신경망의 각 연산 계층간의 연결 관계에 따라 어느 연산 계층들이 융합되는지와 융합 가능한 연산 계층의 수량을 확정할 수 있고, 상기 하나 이상의 연산 계층을 융합할 수 있는 조합을 하나의 목표 연산 동작으로 사용한다. 이와 같이, 신경망의 깊이 방향에서 복수의 연산 계층을 융합하여 하나의 목표 연산 동작으로 사용함으로써, 연산 횟수 및 데이터 판독 횟수를 감소시킬 수 있어, 프로세서의 처리 효율을 더욱 향상시킬 수 있다.

[0109] 예를 들어, 도 9에 제시된 바와 같이, 상기 신경망의 각 연산 계층의 연결 관계에 따라 상기 신경망의 깊이 방향에서 컨벌루션 연산 동작, 풀링 연산 동작 및 활성화 연산 동작이 융합될 수 있는 것을 확정할 수 있다. 여기서, 프로세서는 제 1 메모리의 가용 저장 용량, 및 각 연산 동작의 목표 입력 데이터 용량 등에 따라 융합 수량 역치를 확정할 수 있다. 구체적으로, 제 1 메모리의 가용 저장 용량이 풀링 연산 동작의 목표 입력 데이터 C1-C2, 및 활성화 연산 동작의 목표 입력 데이터 B1-B2를 수용할 수 있을 경우, 상기 융합 수량 역치는 2개로 확정할 수 있고, 상기 풀링 연산 동작 및 활성화 연산 동작은 하나의 목표 연산 동작인 것과 같다. 여기서, 상기 목표 연산 동작의 목표 입력 데이터는 C1-C2 구간 내의 데이터일 수 있다. 기타 실시예에서, 상기 목표 연산 동작은 컨벌루션 연산 동작, 풀링 연산 동작 및 활성화 연산 동작의 융합일 수도 있다. 또는, 상기 활성화 연산 동작 후에 기타 연산 동작이 수행될 때, 상기 목표 연산 동작은 제 1 메모리의 가용 저장 용량에 따라 더 많은 연산 동작을 계속 융합할 수 있다. 예를 들어, 상기 신경망은 N개 연산 계층을 포함할 수 있으며, 프로세서는 상기 제 1 메모리의 가용 저장 용량에 따라 융합 역치가 n 개임을 확정할 수 있고(단, n은 1보다 크거나 같고, n은 N보다 작거나 같음), n개의 연산 계층을 하나의 목표 연산 동작으로 사용할 수 있다. 이것은 단지 설명을 위한 것이며 제한하려는 것은 아니다.

[0110] 또한, 해당 목표 연산 동작이 복수의 연산 동작을 포함하는 경우, 상기 목표 연산 동작의 실행 과정에서의 중간 계산 결과를 제 1 메모리에 저장할 수도 있다. 구체적으로, 상기 방법은 다음 단계를 추가로 포함한다: 상기 목표 연산 동작에서 현재 연산 동작이 출력한 중간 계산 결과가 그 후의 기타 연산 동작의 입력 데이터로 할 필요가 있을 경우, 또는 현재 연산 동작이 출력한 중간 계산 결과가 기타 목표 연산 동작의 입력 데이터로 할 필요가 있을 경우, 프로세서는 상기 현재 연산 동작이 출력한 중간 계산 결과를 제 1 메모리 살에 일시적으로 저장할 수 있다. 구체적으로, 프로세서는 상기 현재 연산 동작이 출력한 중간 결과의 데이터 용량에 따라, 제 1 메모리 상에 상기 현재 연산 동작이 출력한 중간 결과에 저장 주소를 분배할 수 있다. 상기 현재 연산 동작 후의 기타 연산 동작이나 기타 목표 연산 동작이 상기 현재 연산 동작이 출력한 중간 계산 결과를 사용할 필요가 없는 경우, 상기 현재 연산 동작이 출력한 중간 결과가 차지하는 저장 공간을 재분배할 수 있으며, 즉, 상기 현재 연산 동작이 출력한 중간 결과가 차지하던 저장 주소를 기타 데이터에 분배할 수 있다. 예를 들어, 도8에 제시된 바와 같이, 상기 현재 연산 동작OP1에 의해 출력된 중간 계산 결과 Y1이 다음 연산 동작 OP2의 입력 데이터인 경우, 프로세서는 상기 현재 연산 동작에 의해 출력된 중간 결과Y1을 일시적으로 제 1 메모리에 저장할 수 있다. 따라서, 중간 계산 결과 Y1의 판독 횟수가 감소되므로, 프로세서의 처리 효율 및 속도가 향상될 수 있다. 상기 연산 동작OP2 는 중간 계산 결과를 사용할 필요가 없고, 상기 목표 연산 동작 후의 기타 목표 연산 동작에서도 상기 중간 계산 결과 Y1 을 사용할 필요가 없을 경우, 상기 중간 계산 결과Y1이 차지하던 저장 공간을 해제할 수 있고, 상기 중간 계산 결과Y1이 차지하던 저장 주소는 기타 데이터에 분배할 수 있다. 예를 들어, 현재 목표 연산 동작 후의 기타 목표 연산 동작의 목표 출력 데이터를 중간 계산 결과Y1이 차지하던 저장 공간에 저장하여, 제 1 메모리상의 저장 공간 재사용을 실현한다. 기타 예로, 도 9에 제시된 바와 같이, 풀링 연산 동작의 목표 입력 데이터는 C1-C2 구간 내의 데이터이고, 상기 목표 입력 데이터에 대응하는 목표 출력 데이터는 B1-B2 구간의 데이터이다. 그리고, 상기 목표 출력 데이터 B1-B2는 활성화 연산 동작의 목표 입력 데이터이기에, 프로세서는 상기 중간 계산 결과 B1-B2를 제 1 메모리에 일시적으로 저장할 수 있다. 이러한 방식으로, 중간 계산 결과 B1-B2를 판독하는 횟수가 감소되므로, 프로세서의 처리 효율 및 속도가 향상될 수 있다. 활성화 연산 동작이 상기 목표 출력 데이터 B1-B2가 불필요한 경우, 목표 출력 데이터 B1-B2에 의해 점유된 저장 공간은 기타 데이터에 분배되어, 제 1 메모리상의 저장 공간 재사용을 실현한다. 일 실시예에서, 목표 연산 동작의 목표 입력 데이터가 상기 목표 연산 동작에 대응하는 모든 입력 데이터의 일부일 경우, 상기 목표 연산 동작의 각 목표 입력 데이터는 상기 목표 연산 동작의 일부 연산이 완료되도록 구성된다. 상기 목표 연산 동작의 처리 속도를 향상시키고 제 1 메모리의 저장 공간을 최대한 활용하기 위해, 상기 목표 연산 동작에 대응하는 목표 입력 데이터의 수는 하나 이상일 수 있고, 각 목표 입력 데이터는 모든 입력 데이터의 일부, 즉 각각

의 목표 입력 데이터는 모든 입력 데이터의 1개 이상의 입력 데이터 블록을 포함한다. 즉, 하나 이상의 목표 입력 데이터를 함께 제 1 메모리에 로딩할 수 있다. 또한, 상기 목표 입력 데이터의 수량에 따라, 상기 목표 연산 동작은 복수의 서브 목표 연산 동작으로 분할될 수 있고, 선택적으로, 각 서브 목표 연산 동작은 동일한 연산을 수행할 수 있다. 구체적으로, 도 7에 제시된 바와 같이, 상기 방법은 다음 단계를 더 포함한다.

[0111] S500단계: 각 상기 서브 목표 연산 동작에 대응하는 목표 입력 데이터 용량 및 목표 출력 데이터 용량에 따라, 각 서브 목표 연산 동작에 필요한 목표 저장 용량을 확정하고; 각 서브 목표 연산 동작에 필요한 목표 저장 용량은 같거나 다를 수 있다.

[0112] S510단계: 제 1 메모리의 가용 저장 용량 및 현재의 서브 목표 연산 동작에 필요한 목표 저장 용량에 따라, 제 1 메모리의 잔여 저장 용량을 확정한다;

[0113] S520단계: 제 1 메모리의 잔여 저장 용량 및 각 서브 목표 연산 동작에 필요한 목표 저장 용량에 따라, 상기 서브 목표 연산 동작의 수량을 확정한다.

[0114] 선택적으로, 상기 제 1 메모리의 잔여 저장 용량 및 상기 현재 서브 목표 연산 동작 이외의 기타 서브 목표 연산 동작에 필요한 목표 저장 용량에 따라, 상기 제 1 메모리에 서브 목표 연산 동작을 더 수용할 수 있는 수량을 확정할 수 있다. 다음, 상기 현재 서브 목표 연산 동작과 그 외의 기타 서브 목표 연산 동작의 수량에 따라, 상기 서브 목표 연산 동작의 총수량을 확정할 수 있다. 구체적으로, 현재 서브 목표 연산 동작의 목표 입력 데이터 용량과 목표 출력 데이터 용량의 합이 상기 제 1 메모리의 가용 저장 용량보다 작은 경우, 상기 제 1 메모리의 잔여 저장 용량에 따라 하나 이상의 서브 목표 연산 동작을 수행할 수 있는지를 판단할 수 있다. 수행할 수 있을 경우, 프로세서는 하나 이상의 상기 서브 목표 연산 동작에 대응하는 목표 입력 데이터들을 함께 처리할 수 있다. 이와 같이, 복수의 목표 입력 데이터를 함께 처리함으로써, 프로세서의 처리 속도 및 효율을 더욱 향상시킬 수 있다. 도 8에 제지된 바와 같이, 목표 연산 동작(도면에서 좌측의 연산 동작)은 연산 동작 OP1 및 OP2를 포함할 수 있고, 프로세서는 상기 목표 연산 동작의 현재 서브 목표 연산 동작의 목표 입력 데이터 X11, X21 및 Y3의 데이터 용량을 확정하고,상기 현재 서브 목표 연산 동작의 목표 출력 데이터 Y1, Z1의 데이터 용량을 확정하는 것에 의하고, 현재 서브 목표 연산 동작의 목표 입력 데이터와 목표 출력 데이터의 합에 의해, 현재 서브 목표 연산 동작에 필요한 목표 저장 용량을 확정할 수 있다. 상기 현재 서브 목표 연산 동작에 필요한 목표 저장 용량이 상기 제 1 메모리의 가용 저장 용량보다 작은 경우, 상기 제 1 메모리의 잔여 저장 용량은 계산을 통해 얻을 수 있다. 상기 제 1 메모리의 잔여 저장 용량은 제 1 메모리의 가용 저장 용량에서 상기 현재 서브 목표 연산 동작의 목표 저장 용량을 뺀 것과 동일하다. 다음, 프로세서는 상기 제 1 메모리의 잔여 저장 용량에 따라 서브 목표 연산 동작의 수량을 확정할 수 있다. 구체적으로, 상기 제 1 메모리의 잔여 저장 용량이 기타 서브 목표 연산 동작의 목표 입력 데이터 X12, X22 ,Y4, 및 연산 동작 OP1이 출력한 중간 계산 결과 Y2, 및 연산 동작 OP2가 출력한 목표 출력 데이터Z2를 추가로 수용할 수 있을 경우, 상기 목표 연산 동작의 수량는 2개로 확정할 수 있고, 상기 서브 입력 데이터 X21, X22 및 Y4를 그 중의 1개의 목표 연산 동작의 목표 입력 데이터로 사용할 수 있다. 이와 같이, 상기 처리하고자 하는 연산의 가로 방향에서 동일한 목표 연산 동작의 복수의 목표 입력 데이터를 동시에 로딩함으로써, 프로세서는 복수의 목표 입력 데이터를 병렬 처리하게 될 수 있어, 프로세서의 처리 속도 및 효율을 더욱 향상시킬 수 있다.또한, 상기 제 1 메모리의 잔여 저장 용량이 기타 서브 목표 연산 동작의 목표 입력 데이터 X12, X22, Y4, 및 연산 동작 OP1이 출력한 중간 계산 결과 Y2, 및 연산 동작 OP2가 출력한 목표 출력 데이터Z2를 추가로 수용할 수 있을 뿐만 아니라, 상기 제 1 메모리의 잔여 저장 용량이 연산 동작 OP3의 출력 데이터 Y도 수용할 수 있을 경우, 연산 동작 OP1, OP2, OP3을 융합할 수 있어, 1회의 연산 수행으로 계산 결과Y 를 얻을 수 있다. 기타 예로서, 도 9에 제시된 바와 같이, 상기 처리하고자 하는 연산은 신경망 연산 등과 같은 연산이고, 상기 처리하고자 하는 연산은 컨벌루션 계층, 풀링 계층 및 활성화 계층을 포함할 수 있으며, 전술한 각 연산 계층의 실행 순서는 차례로 컨벌루션 연산 동작-풀링 연산 동작-활성화 연산 동작이다. 상기 목표 연산 동작이 활성화 연산 동작인 경우, 프로세서는 제 1 메모리의 저장 용량에 따라, 현재 서브 목표 연산 동작의 목표 입력 데이터를 획득할 수 있고, 상기 현재 서브 목표 연산 동작의 목표 입력 데이터는 풀링 계층 상 B1-B2구간 내의 입력 데이터일 수 있다. 상기 현재 서브 목표 연산 동작의 목표 출력 데이터는 A1이다. 상기 현재 서브 목표 연산 동작의 목표 입력 데이터의 데이터 용량 B1-B2와 그에 대응하는 목표 출력 데이터의 데이터 용량의 합이 상기 제 1 메모리의 저장 용량보다 작은 경우, 즉 상기 현재 서브 목표 연산 동작에 필요한 목표 저장 용량이 제 1 메모리의 저장 용량보다 작은 경우, 프로세서는 진일보 상기 제 1 메모리의 잔여 저장 용량에 따라 상기 서브 목표 연산 동작의 수량을 확정할 수 있다. 예를 들어, 프로세서는 상기 제 1 메모리의 잔여 저장 용량에 따라, 제 1 메모리의 잔여 저장 용량이 활성화 연산 A1-A2 구간의 연산량을 만족시킬 수 있다고 확정되면, 서브 목표 연산 동작의 대상의 수량은 2개이고, 목표 입력 데이터 B2-

B3 구간 내의 데이터를 상기 활성화 연산 동작의 1개 목표 연산 동작에 대응하는 목표 입력 데이터로 사용할 수 있다. 또한, 1개 이상의 서브 목표 연산 동작의 목표 입력 데이터가 교차한다면, 현재 서브 목표 연산 동작의 목표 입력 데이터와 기타 서브 목표 연산 동작의 목표 입력 데이터의 교집합을 확정하고, 상기 교집합을 제 1 메모리 상에 임시 저장한다. 즉, 현재 서브 목표 연산 동작의 목표 입력 데이터의 일부 또는 전부가 기타 서브 목표 연산 동작의 목표 입력 데이터이기도 한 경우, 상기 교집합을 제 1 메모리에 임시 저장하여 상기 부분 데이터의 복수 판독 동작을 피할 수 있으므로, 프로세서의 처리 속도 및 효율을 향상시킬 수 있다. 예를 들어, 도 9에 제시된 바와 같이, 목표 연산 동작이 풀링 연산 동작 인 경우, 상기 목표 연산 동작의 서브 목표 연산 동작의 수량은 2개이고, 대응하여 상기 풀링 연산 동작에 대응하는 목표 입력 데이터는 2개일 수 있고, 그 중 하나의 목표 입력 데이터는 C1-C2이고, 기타 하나의 목표 입력 데이터는 C3-C4이다. 상기 목표 입력 데이터 C1-C2에 대응하는 목표 출력 데이터는 B1-B2이고, 상기 목표 입력 데이터 C3-C4에 대응하는 목표 출력 데이터는 B2-B3이다. 첨부된 도면에 따리, 입력 데이터 C3-C2 구간의 데이터는 목표 입력 데이터C1-C2의 일부이고, 한편 목표 입력 데이터 C3-C4의 일부이기도 하고, 즉 두 개의 목표 입력 데이터에는 교집합 C3-C2 가 존재한다. 여기서, 데이터의 판독 횟수를 줄이기 위해, 상기 목표 입력 데이터 C1-C2가 대응하는 풀링 연산을 완료된 후, 입력 데이터 C3-C2가 여전히 제 1 메모리에 저장되어, 이 부분 데이터의 복수의 판독 동작를 피함으로써, 프로세서의 처리 효율 및 속도를 향상시킬 수 있다.

[0115] 일 실시예에서, 상기 방법에 다음 단계를 추가로 포함한다 :

[0116] 목표 연산 동작 후의 연산 동작과 목표 연산 동작간의 연산 간격이 설정된 범위 내에 있을 경우, 목표 출력 데이터가 제 1 메모리에 저장되어, 목표 출력 데이터의 판독 횟수를 감소시킨다.

[0117] 구체적으로, 목표 연산 동작 후의 연산 동작과 상기 목표 연산 동작간의 연산 간격이 설정된 범위 내에 있을 경우, 예를 들어, 목표 연산 동작과 그 후의 기타 연산 동작 간의 연산 간격이 3~5개 연산 동작이면, 상기 목표 출력 데이터는 제 1 메모리에 저장되어, 목표 출력 데이터가 판독되는 횟수를 감소시킬 수 있다. 목표 연산 동작과 그 후의 기타 연산 동작 간의 연산 간격이 설정된 범위를 초과하면, 상기 목표 출력 데이터가 긴 시간 동안 제 1 메모리의 저장 공강을 점유하는 것을 피하기 위해, 상기 목표 출력 데이터를 제 1 메모리로부터 제 2 메모리로 전송한다.

[0118] 일 실시예에서, 상기 방법은 추가로 다음 단계를 더 포함한다 :

[0119] 프로세서는 목표 연산 동작의 목표 입력 데이터의 데이터 용량에 따라, 제 1 메모리 상에 목표 입력 데이터의 저장 주소를 확정할 수 있고; 목표 연산 동작의 목표 출력 데이터의 데이터 용량에 따라, 제 1 메모리 상에 목표 출력 데이터의 저장 주소를 확정할 수 있다. 구체적으로는, 프로세서는 목표 연산 동작의 목표 입력 데이터의 데이터 용량에 따라, 제 1 메모리에서 상기 목표 입력 데이터에게 그 데이터 용량에 매칭되는 저장 공간을 분배할 수 있으며, 상기 저장 공간의 저장 주소를 상기 목표 입력 데이터에 분배할 수 있다. 따라서, 실제 연산 과정에서, 목표 입력 데이터는 제 1 메모리 상의 지정된 저장 공간에 로딩될 수 있다. 마찬가지로, 프로세서는 목표 연산 동작의 목표 출력 데이터의 데이터 용량에 따라, 제 1 메모리 상에 상기 목표 출력 데이터에게 그 데이터 용량에 매칭되는 저장 공간을 분배할 수 있으며, 상기 저장 공간의 저장 주소를 상기 목표 출력 데이터에 분배할 수 있다. 따라서, 실제 연산 과정에서, 목표 출력 데이터는 제 1 메모리 상의 지정된 저장 공간에 저장할 수 있다.

[0120] 일 실시예에서, 상기 방법은 다음과 같은 단계를 더 포함한다:

[0121] 목표 연산 동작의 목표 입력 데이터가 계속 사용되지 않을 경우, 프로세서는 목표 입력 데이터의 저장 공간의 전부 또는 일부를 목표 연산 동작의 목표 출력 데이터에 분배할 수 있다. 이러한 방식으로, 동일한 저장 공간 여러번 재사용하여, 상기 제 1 메모리의 공간 이용율을 향상시킬 수 있다. 선택적으로, 프로세서는 각 목표 연산 동작의 목표 입력 데이터의 저장 주소, 목표 출력 데이터의 저장 주소, 중간 계산 결과의 저장 주소, 제 1 메모리 상의 각 저장 공간의 업데이트 규칙 등을 기록할 수 있고, 상기 데이터에 대응하는 저장 주소에 따라 상기 처리하고자 하는 연산에 대응하는 저장 분배 규칙을 획득한다. 프로세서가 상기 처리하고자 하는 연산을 수행할 필요가 있을 경우, 프로세서는 상기 처리하고자 하는 연산에 대응하는 저장 분배 규칙을 획득하고, 상기 저장 분배 규칙에 따라 연산 과정에서 각종 데이터의 리드-라이트 동작 및 저장 위치 등을 확정할 수 있다. 일 실시예에서, 전술한 데이터 전처리 방법은 도 2 ~도 4에 명시된 컴퓨터 설비에도 적용될 수 있다. 여기서, 설정된 연산 분배 규칙에 따라, 상기 목표 연산 동작의 일부는 메인 프로세싱 회로에 의해 실행될 필요가 있고, 상기 목표 연산 동작의 기타 일부는 슬레이브 프로세싱 회로에 의해 실행될 필요가 있다. 따라서, 복수의 제 1 메모리는 메인 메모리 및 슬레이브 메모리를 포함할 수 있으며, 여기서 메인 메모리는 메인 프로세싱 회로에 가깝

게 배치되고, 또한 상기 메인 메모리는 메인 프로세싱 회로의 온-칩 메모리일 수도 있다. 상기 슬레이브 메모리는 슬레이브 프로세싱 회로에 근접하여 배치되며, 상기 슬레이브 메모리는 슬레이브 프로세싱 회로의 온-칩 메모리일 수도 있다. 여기서, 상기 목표 연산 동작에 대응하는 목표 입력 데이터의 일부는 상기 메인 메모리에 로딩되어 메인 프로세싱 회로에 의해 실행될 필요가 있고, 상기 목표 연산 동작에 대응하는 목표 입력 데이터의 기타 일부는 하나 이상의 슬레이브 메모리에 로딩되어, 각 슬레이브 메모리에 대응하는 슬레이브 회로에 의해 실행될 필요가 있다.

[0122] 구체적으로, 도 10에 제시된 바와 같이, 도 2 ~도 4에 제시된 컴퓨터 설비가 전술한 데이터 전처리 방법을 실행할 때, 하기 단계를 포함한다.

[0123] S600단계: 메인 메모리의 가용 저장 용량, 슬레이브 메모리의 가용 저장 용량 및 목표 연산 동작을 얻는다. 구체적으로, 프로세서는 상기 메인 메모리의 구성 정보(예: 모델번호 등 정보)에 따라, 상기 메인 메모리의 총 저장 용량을 획득할 수 있다. 또한, 프로세서는 상기 메인 메모리의 총 저장 용량 및 상기 메인 메모리 상의 점유된 저장 용량에 따라, 상기 메인 메모리의 가용 저장 용량을 얻을 수 있다. 유사하게, 프로세서는 상기 슬레이브 메모리의 구성 정보에 따라, 슬레이브 메모리의 총 저장 용량을 획득하고, 상기 슬레이브 메모리의 총 저장 용량 및 상기 슬레이브 메모리 상의 점유된 저장 용량에 따라, 상기 슬레이브 메모리의 가용 저장 용량을 얻을 수 있다. 선택적으로, 상기 프로세서의 메인 프로세싱 회로는 상기 메인 메모리의 가용 저장 용량을 획득할 수 있고, 각 슬레이브 프로세싱 회로는 대응하는 슬레이브 메모리의 가용 저장 용량을 획득할 수 있고, 대응하는 슬레이브 메모리의 가용 저장 용량을 메인 프로세싱 회로로 전송할 수 있다. 한편, 프로세서의 제어기 유닛은 처리하고자 하는 연산을 획득할 수 있고, 상기 처리하고자 하는 연산의 분석 결과 등과 같은 데이터를 메인 프로세싱 회로에 전송한다. 상기 메인 프로세싱 회로는 상기 처리하고자 하는 연산, 메인 메모리의 가용 저장 용량 및 슬레이브 메모리의 가용 저장 용량에 따라 목표 연산 동작을 확정할 수 있다. 선택적으로, 상기 처리하고자 하는 연산은 덧셈 동작, 뺄셈 동작, 곱셈 동작, 나눗셈 동작, 컨벌루션 동작, 풀링(Pooling) 동작 및 활성화(예 : Relu) 동작 등 연산 동작을 포함할 수 있으며, 여기서 구체적인 제한하지 않는다. 상기 목표 연산 동작은 처리하고자 하는 연산에서 하나 이상의 연산 동작의 조합일 수 있다.

[0124] S700단계: 상기 메인 메모리의 가용 저장 용량, 상기 슬레이브 메모리의 가용 저장 용량 및 상기 목표 연산 동작에 따라, 상기 목표 연산 동작에 대응하는 목표 입력 데이터를 확정한다. 여기서, 상기 목표 입력 데이터는 상기 목표 연산 동작에 대응하는 모든 입력 데이터의 일부 또는 전부이다.

[0125] 구체적으로는, 프로세서의 메인 프로세싱 회로는 상기 목표 연산 동작에 따라, 상기 목표 연산 동작이 완료되기 위해 필요한 전부 입력 데이터 및 상기 전부 입력 데이터의 데이터 용량(즉, 상기 전부 입력 데이터에 수요되는 저장 공간의 크기)을 확정할 수 있다. 또한, 메인 프로세싱 회로는 메인 메모리의 가용 저장 용량, 각 슬레이브 메모리의 가용 저장 용량 및 상기 목표 연산 동작의 전부 입력 데이터의 데이터 용량에 따라, 상기 목표 연산 동작에 대응하는 목표 입력 데이터 및 데이터 용량을 확정할 수 있다.

[0126] S800단계: 상기 목표 연산 동작 및 상기 목표 입력 데이터에 따라, 상기 목표 연산 동작에 대응하는 목표 출력 데이터를 확정한다. 구체적으로, 처리하고자 하는 연산의 계산량은 정적 분석 가능하기에, 프로세서의 메인 프로세싱 회로는 상기 목표 연산 동작의 목표 입력 데이터 및 상기 목표 연산 동작에 따라, 상기 목표 연산 동작의 목표 출력 데이터 및 상기 목표 출력 데이터의 데이터 용량 등 정보를 얻을 수 있고, 즉 프로세서의 메인 프로세싱 회로는 상기 목표 연산 동작의 목표 출력 데이터에 필요한 저장 공간을 얻을 수 있다. S900: 상기 목표 연산 동작의 목표 출력 데이터가 상기 목표 연산 동작 후의 기타 연산 동작의 입력 데이터인 경우, 상기 목표 출력 데이터는 상기 메인 메모리에 대응하여 저장된다. 구체적으로, 메인 프로세싱 회로는 설정된 연산 분배 규칙에 따라, 상기 목표 연산 동작에 대응하는 목표 입력 데이터를 메인 메모리 및 슬레이브 메모리에 분배하여, 메인 프로세싱 회로와 슬레이브 프로세싱 회로가 협력하여 목표 연산 동작을 수행할 수 있도록 한다. 상기 목표 연산 동작 실행 과정 중에, 슬레이브 프로세싱 회로는 자체의 슬레이브 메모리 위의 목표 입력 데이터를 처리하여 중간 계산 결과를 얻을 수 있다. 또한, 슬레이브 프로세싱 회로는 상기 중간 계산 결과를 메인 프로세싱 회로로 전송할 수 있다. 메인 프로세싱 회로는 자체의 메인 메모리 상의 목표 입력 데이터를 처리하고, 각 슬레이브 프로세싱 회로로부터 전송된 중간 계산 결과와 결합하여, 상기 목표 연산 동작의 목표 출력 데이터를 얻을 수 있다. 상기 목표 연산 동작의 목표 출력 데이터가 그 후의 기타 연산 동작의 입력 데이터인 경우, 상기 목표 출력 데이터가 메인 메모리 상에 저장될 수 있어, 데이터의 판독 횟수를 감소시키고 프로세서의 계산 속도를 향상시킬 수 있다.

[0127] 일 실시예에서, 도 11에 제시된 바와 같이, 상기 S700단계는,

[0128] 상기 메인 메모리의 가용 저장 용량을 각 상기 슬레이브 메모리의 가용 저장 용량과 비교하여, 최소의 가용 저장 용량을 제 1 메모리의 가용 저장 용량으로서 사용하는 S710단계;

[0129] 제 1 메모리의 가용 저장 용량 및 목표 연산 동작에 따라 목표 연산 동작에 대응하는 목표 입력 데이터를 확정하는 S720단계를 포함한다. 구체적으로, 상기 목표 연산 동작은 메인 프로세싱 회로와 슬레이브 프로세싱 회로를 연동시켜 완성할 필요가 있기 때문에, 메인 메모리와 슬레이브 메모리가 상기 목표 연산 동작의 목표 입력 데이터의 점유 공간을 동시에 만족시키는 것이 보증될 필요가 있다. 즉, 상기 목표 입력 데이터의 데이터 용량은 상기 메인 메모리의 가용 저장 용량보다 작으며, 상기 목표 입력 데이터의 데이터 용량은 상기 슬레이브 메모리의 가용 저장 용량보다 작다. 따라서, 메인 메모리의 가용 저장 용량과 각 슬레이브 메모리의 가용 저장 용량을 비교하여, 메인 메모리 및 각 슬레이브 메모리 중 최소 가용 저장 용량을 상기 프로세서의 제 1 메모리의 가용 저장 용량으로서 사용한다. 그리고, 메인 프로세싱 회로는 상기 제 1 메모리의 가용 저장 용량 및 목표 연산 동작에 따라, 상기 목표 연산 동작에 대응하는 목표 입력 데이터를 확정할 수 있다. 일 실시예에서, 메인 프로세싱 회로는 설정된 연산 분배 규칙에 따라 상기 목표 입력 데이터를 복수의 데이터 블록로 분할할 수 있고, 각 데이터 블록에 대응하는 프로세싱 회로를 확정할 수 있다. 여기서, 상기 목표 입력 데이터에서, 메인 프로세싱 회로에 의해 처리되는 데이터 블록은 제 1 목표 입력 데이터로 표기할 수 있다. 상기 목표 입력 데이터에서, 슬레이브 프로세싱 회로에 의해 처리되는 데이터 블록은 제 2 목표 입력 데이터로 표기할 수 있다. 또한, 각 슬레이브 프로세싱 회로에 대응하는 제 2 목표 입력 데이터의 데이터 용량은 다를 수 있으며, 구체적으로 연산 분배 규칙에 의해 결정된다. 구체적으로, 상기 방법은 다음 단계를 더 포함한다 :설정된 연산 분배 규칙에 따라, 상기 메인 메모리에 대응하는 제 1 목표 입력 데이터 및 상기 슬레이브 메모리 각각에 대응하는 제 2 목표 입력 데이터를 확정한다. 구체적으로, 메인 프로세싱 회로는 설정된 연산 분배 규칙에 따라, 상기 목표 연산 동작의 어떤 목표 입력 데이터가 메인 프로세싱 회로에 의해 처리되고, 상기 목표 연산 동작의 어떤 목표 입력 데이터가 각 슬레이브 프로세싱 회로에 의해 처리되는지를 확정할 수 있다. 예를 들면, 도 9에 제시된 바와 같이, 상기 현재 연산 동작은 풀링 연산 동작으로서, 풀링 계층 상 B1-B2 구간의 연산을 완료할 필요가 있을 경우, 상기 목표 연산 동작에 필요한 목표 입력 데이터는 C1-C2이다. 메인 프로세서는 설정된 연산 분배 규칙에 따라, 입력 데이터 C1-C3을 제 2 목표 입력 데이터로 사용하고, 상기 제 2 목표 입력 데이터 C1-C3을 슬레이브 메모리에 저장할 수 있다. 입력 데이터 C3-C2는 제 1 목표 입력 데이터로 사용하고, 상기 제 1 목표 입력 데이터 C3-C2는 메인 메모리 내에 저장할 수 있다.

[0130] 또한, 상기 방법은 다음 단계를 더 포함할 수 있다: 프로세서는 메인 메모리의 가용 저장 용량 및 상기 제 1 목표 입력 데이터의 데이터 용량에 따라, 메인 메모리 상의 제 1 목표 입력 데이터의 저장 주소를 확정할 수 있다. 구체적으로, 메인 프로세싱 회로는 메인 메모리의 가용 저장 용량 및 제 1 목표 입력 데이터의 데이터 용량에 따라, 메인 메모리 상의 상기 제 1 목표 입력 데이터의 저장 주소를 확정할 수 있다. 또한, 상기 메인 프로세싱 회로는 상기 제 1 목표 입력 데이터 용량 및 목표 연산 동작에 따라, 상기 제 1 목표 입력 데이터에 대응하는 제 1 목표 출력 데이터 및 그 데이터 용량을 확정할 수 있고, 메인 메모리 상의 상기 제 1 목표 출력 데이터의 저장 주소도 확정할 수 있다. 프로세서는 슬레이브 메모리의 가용 저장 용량 및 제 2 목표 입력 데이터의 데이터 용량에 따라, 슬레이브 메모리 상의 제 2 목표 입력 데이터의 저장 주소를 확정할 수 있다. 구체적으로, 메인 프로세싱 회로는 각 슬레이브 프로세싱 회로의 가용 저장 용량 및 그에 대응하는 제 2 목표 입력 데이터의 데이터 용량에 따라, 각 제 2 목표 입력 데이터가 그에 대응하는 슬레이브 메모리에서의 저장 주소를 확정할 수 있다. 또한, 상기 메인 프로세싱 회로는 각 제 2 목표 입력 데이터 용량 및 목표 연산 동작에 따라, 각 제 2 목표 입력 데이터에 대응하는 제 2 목표 출력 데이터 및 그 데이터 용량을 확정할 수 있고, 각 제 2 목표 출력 데이터가 그에 대응하는 슬레이브 메모리에서의 저장 주소도 확정할 수 있다.

[0131] 또한, 각 슬레이브 프로세싱 회로는 얻은 제 2 목표 출력 데이터를 메인 프로세싱 회로에 전송하고, 메인 프로세싱 회로는 추가로 메인 메모리 상의 상기 제 2 목표 출력 데이터 각각의 저장 주소를 확정할 수 있다. 일 실시예에서, 상기 슬레이브 프로세싱 회로에 의해 수행된 기타 연산 동작은 그에 대응한 상기 제 2 목표 출력 데이터를 계속 필요할 경우, 상기 제 2 목표 출력 데이터는 상기 슬레이브 프로세싱 회로에 대응하는 슬레이브 메모리 상에 일시적으로 저장할 수 있다. 이러한 방식으로, 메인 메모리와 슬레이브 메모리 사이의 데이터 판독 동작이 감소될 수 있고, 상기 프로세서의 연산 속도가 더 향상될 수 있다. 일 실시예에서, 전술한 목표 연산 동작은 하나 이상의 연산 동작을 포함하며. 즉, 상기 목표 연산 동작은 하나 이상의 연산 동작의 조합이다. 일반적으로, 상기 목표 연산 동작에 포함되는 각 연산 동작은 서로 기타 연산을 구현하기 위한 상이한 기타 연산 동작으로 한다. 이 경우, 프로세서의 메인 프로세싱 회로는 제 1 메모리의 가용 저장 용량에 따라, 각 연산 동작에 대응하는 서브 목표 입력 데이터를 확정할 수 있고, 각 연산 동작에 대응하는 서브 목표 입력 데이터에 따라

상기 목표 연산 동작에 대응하는 목표 입력 데이터를 확정할 수 있다. 구체적으로, 상기 목표 입력 데이터를 확정하는 과정은 전술한 방법의 S210 ~ S230단계와 일치하며, 상세한 내용은 상기 설명을 참조할 수 있고, 상세한 설명은 여기서 생략한다. 또한, 상기 목표 연산 동작이 복수의 연산 동작을 포함하는 경우, 상기 하나 이상의 연산 동작은 제 1 목표 연산 동작과 제 2 목표 연산 동작으로 분할할 수 있다. 메인 프로세싱 회로는 설정된 연산 분배 규칙에 따라, 상기 목표 연산 동작 중의 제 1 목표 연산 동작을 메인 프로세싱 회로에 할당하고, 상기 목표 연산 동작 중의 제 2 목표 연산 동작을 슬레이브 프로세싱 회로에 할당할 수 있다. 이에 따라, 메인 프로세싱 회로는 제 1 목표 연산 동작에 필요한 입력 데이터를 메인 메모리에 저장하고, 상기 제 2 목표 연산 동작 각각에 필요한 입력 데이터를 그에 대응하는 슬레이브 메모리에 저장할 수 있다. 예를 들어, 도 9에 제시되 바와 같이, 제 1 메모리의 가용 저장 용량이 풀링 연산 동작의 목표 입력 데이터 C1-C2, 및 활성화 연산 동작의 목표 입력 데이터 B1-B2를 수용할 수 있을 경우, 상기 풀링 연산 동작과 활성화 연산 동작은 하나의 목표 연산 동작인 것과 같다. 이 경우, 상기 목표 연산 동작의 목표 입력 데이터는 C1-C2 구간 내의 데이터일 수 있다. 여기서, 메인 프로세싱 회로는 설정된 연산 분배 규칙에 따라, 활성화 연산 동작을 제 1 목표 연산 동작으로 사용하고 메인 프로세싱 회로에 할당할 수 있고, 풀링 연산 동작을 제 2 목표 연산 동작으로 사용하고 슬레이브 프로세싱 회로에 할당할 수 있다. 이에 따라, 풀링 연산 동작에 필요한 입력 데이터 C1-C2는 슬레이브 메모리에 로딩될 수 있고, 활성화 연산 동작에 필요한 입력 데이터 B1-B2는 메인 메모리에 로딩될 수 있다. 상기 풀링 연산 동작과 활성화 연산 동작 간에 의존관계가 있기 때문에, 상기 풀링 연산 동작이 완료된 후에 활성화 연산 동작에 필요한 입력 데이터(B1-B2)를 메모리로부터 메인 메모리에 로딩할 수 있다.일 실시예에서, 목표 연산 동작의 목표 입력 데이터가 상기 목표 연산 동작에 대응하는 모든 입력 데이터의 일부일 경우, 상기 목표 연산 동작의 각 목표 입력 데이터는 단지 상기 목표 연산 동작의 일부 연산만 완료하기 위한 것이다. 목표 연산 동작의 처리 속도를 향상시키고 제 1 메모리의 저장 공간을 최대한 활용하기 위해, 상기 목표 연산 동작에 대응하는 목표 입력 데이터의 수량은 하나 이상일 수 있고, 각 목표 입력 데이터는 모든 입력 데이터의 일부, 즉 각 목표 입력 데이터는 모든 입력 데이터의 하나 이상의 입력 데이터 블록을 포함한다. 즉, 하나 이상의 목표 입력 데이터가 함께 제 1 메모리 상에 로딩될 수 있다. 또한, 상기 목표 입력 데이터의 수량에 따라, 상기 목표 연산 동작을 복수의 서브 목표 연산 동작으로 분할할 수 있고, 선택적으로, 각 서브 목표 연산 동작은 동일한 연산을 실현할 수 있다.

[0132] 상기 메인 프로세싱 회로는 제 1 메모리의 가용 저장 용량 및 각 목표 연산 동작에 필요한 목표 저장 용량에 따라, 목표 연산 동작의 수량을 확정할 수 있으므로, 하나 이상의 목표 연산 동작의 목표 입력 데이터를 함께 제 1 메모리 상에 로딩할 수 있다. 구체적으로는, 상기 목표 연산 동작의 수량을 확정하는 과정은 전술한 방법의 S500~S520단계와 동일하며, 상세한 내용은 상기 설명을 참조할 수 있으므로, 상세한 설명은 생략한다.

[0133] 도 5~도 7의 흐름도 및 도 10 ~도 11의 각 단계는 화살표에 따라 순차적으로 표시되었지만, 이러한 단계들이 반드시 화살표가 표시한 순서로 수행되는 것은 아니라는 점에 유의해야 한다. 본 문에 명확하게 언급한 경우를 제외하고, 이러한 단계의 수행은 엄격한 순서 제한이 없으며, 이러한 단계들은 기타 순서로 수행될 수도 있다. 또한, 도 5~7 및 도 10 ~도 11중의 적어도 일부 단계는 복수의 서브 단계 또는 복수의 절차를 포함할 수 있으며, 이러한 서브 단계 또는 절차는 반드시 동시에 수행되는 것은 아니며, 상이한 시간에 수행될 수도 있고, 해당 서브 단계 또는 잘차의 수행 순서는 반드시 순차적으로 수행되는 것은 아니며, 기타 단계 또는 기타 단계의 서브 단계 또는 절차의 적어도 일부와 번갈아서 또는 교대로 수행될 수도 있다. 당업자는 전술한 실시예를 구현하기 위한 방법의 전부 또는 일부 프로세스는, 컴퓨터 프로그램에 의해 관련 하드웨어를 명령하는 것으로 완성될 수 있음을 이해할 것이고, 상기 컴퓨터 프로그램은 비 휘발성 컴퓨터 판독 가능 저장 매체에 저장할 수 있고, 상기 컴퓨터 프로그램이 실행될 경우 전술한 각 방법의 실시예의 프로세스를 포함할 수 있다. 여기서 본 개시에 제공된 다양한 실시예에서 사용된 메모리, 저장, 데이터베이스 또는 기타 매체에 대한 모든 인용는 비 휘발성 및/또는 휘발성 메모리를 모두 포함할 수 있다. 비 휘발성 메모리는 판독 전용 메모리(ROM), 프로그래머블 ROM(PROM), 전기적 프로그래머블 ROM(EPROM), 전기적 소거 가능 프로그래머블 ROM(EEPROM) 또는 플래시 메모리를 포함할 수 있다. 휘발성 메모리는 랜덤 액세스 메모리(RAM) 또는 외부 고속 버퍼 메모리를 포함할 수 있다. 설명으로 제한적인 것은 아니며, RAM은 여러 형태로 얻을 수 있으며, 예로 정적 RAM (SRAM), 동적 RAM(DRAM), 동기식 DRAM(SDRAM), 더블 데이터 레이트 SDRAM (DDRSDRAM), 향상형 SDRAM(ESDRAM), 동기 링크(Synchlink) DRAM (SLDRAM), 메모리 버스(Rambus) 다이렉트 RAM(RDRAM), 다이렉트 메모리 버스 DRAM(DRDRAM) 및 메모리 버스 DRAM(RDRAM) 등이다.

[0134] 일 실시예에서, 도 12에 제시된 바와 같이, 본 개시의 실시예는 획득 모듈(410), 입력 확정 모듈(420), 출력 확정 모듈(430) 및 저장 할당 모듈(440)을 포함할 수 있는 데이터 전처리 장치를 제공한다. 여기서, 획득 모듈(410)은 제 1 메모리의 가용 저장 용량 및 목표 연산 동작을 획득하도록 구성되며; 입력 확정 모듈(420)은 상기

목표 연산 동작 및 상기 제 1 메모리의 가용 저장 용량에 따라, 상기 목표 연산 동작에 대응하는 목표 입력 데이터를 확정하도록 구성되며; 출력 확정 모듈(430)은 상기 목표 연산 동작 및 상기 목표 입력 데이터에 따라, 상기 목표 연산 동작의 목표 출력 데이터를 확정하도록 구성되며; 저장 할당 모듈(440)은 상기 목표 연산 동작의 목표 출력 데이터가 상기 목표 연산 동작 후의 기타 연산 동작의 입력 데이터인 경우, 상기 목표 연산 동작의 목표 출력 데이터를 상기 제 1 메모리에 저장하도록 구성되며, 상기 제 1 메모리는 프로세서에 인접하여 배치된다. 선택적으로, 상기 목표 연산 동작은 하나 이상의 연산 동작을 포함하고, 각 연산 동작은 그에 대응하는 서브 목표 입력 데이터가 있다. 도 13에 제시된 바와 같이, 상기 입력 확정 모듈(420)은 융합 확정 유닛(421) 및 입력 확정 유닛(422)를 더 포함한다. 여기서, 융합 확정 유닛(421)은 상기 제 1 메모리의 가용 저장 용량 및 상기 처리하고자 하는 연산에서 각 연산 동작의 융합 속성에 따라, 융합 가능한 연산 동작의 수량을 확정하여 융합 수량 역치를 얻도록 구성된다. 입력 확정 유닛(422)은, 수량이 선정된 상기 융합 가능한 연산 동작의 조합을 상기 목표 연산 동작으로 사용하고, 상기 선정된 수량은 상기 융합 수량 역치 이하이며, 상기 수량이 선정된 융합 가능한 연산 동작 각각에 대응하는 서브 목표 입력 데이터를 상기 목표 연산 동작에 대응하는 목표 입력 데이터로 사용하도록 구성된다.

[0135] 선택적으로, 상기 처리하고자 하는 연산은 복수의 연산 계층을 포함하는 신경망 연산이며, 상기 연산 계층 각각은 하나의 상기 연산 동작을 나타내며; 상기 융합 확정 유닛(421)은 상기 신경망 연산의 각 연산 계층의 연결 관계에 따라 상기 연산 동작 각각의 융합 속성을 확정하도록 구성된다.

[0136] 선택적으로, 상기 저장 할당 모듈(440)은 상기 목표 연산 동작에서 현재 연산 동작에 의해 출력된 중간 계산 결과가 상기 목표 연산 동작 중 기타 연산 동작의 입력 데이터로 하는 경우, 또는 상기 현재 연산 동작에 의해 출력된 중간 계산 결과가 기타 목표 연산 동작의 입력 데이터로 하는 경우, 상기 현재 연산 동작에 의해 출력된 중간 계산 결과를 제 1 메모리 상에 저장하게 되거나, 혹은 상기 현재 연산 동작에 의해 출 된 중간 계산 결과를 제 1 메모리와 제 2 메모리 상에 저장하게 되도록 구성된다. 선택적으로, 상기 목표 연산 동작은 하나 이상의 서브 목표 연산 동작을 포함하고, 상기 서브 목표 연산 동작 각각은 하나의 상기 목표 입력 데이터에 대응하며; 여기서 상기 목표 연산 동작에 대응하는 모든 입력 데이터는 복수의 입력 데이터 블록을 포함하고, 상기 목표 연산 동작에 대응하는 목표 입력 데이터의 수량이 하나 이상이며, 상기 목표 입력 데이터 각각은 하나 이상의 상기 입력 데이터 블록을 포함하고; 입력 확정 모듈(420)은 상기 각 서브 목표 연산 동작 각각의 목표 입력 데이터의 데이터 용량 및 목표 출력 데이터의 데이터 용량에 따라, 상기 서브 목표 연산 동작 각각에 필요한 목표 저장 용량을 확정하고; 상기 제 1 메모리의 가용 저장 용량 및 현재의 서브 목표 연산 동작에 필요한 목표 저장 용량에 따라 상기 제 1 메모리의 잔여 저장 용량을 확정하고; 상기 제 1 메모리의 잔여 저장 용량 및 상기 현재의 서브 목표 연산 동작 이외의 기타 서브 목표 연산 동작에 필요한 목표 저장 용량에 따라, 상기 서브 목표 연산 동작의 수량을 확정하도록 구성된다.

[0137] 선택적으로, 상기 저장 할당 모듈(440)은 하나 이상의 서브 목표 연산 동작의 목표 입력 데이터에 교집합이 존재할 경우, 상기 하나 이상의 서브 목표 연산 동작의 목표 입력 데이터 사이의 교집합을 상기 제 1 메모리 상에 저장하도록 구성된다.

[0138] 선택적으로, 상기 저장 할당 모듈(440)은 상기 목표 연산 동작의 목표 입력 데이터의 데이터 용량에 따라, 상기 제 1 메모리 상에 있는 상기 목표 입력 데이터의 저장 주소를 확정하고; 상기 목표 연산 동작의 목표 출력 데이터의 데이터 용량에 따라, 상기 제 1 메모리 상에 있는 상기 목표 출력 데이터의 저장 주소를 확정하며; 상기 목표 연산 동작 후의 기타 연산 동작이 상기 목표 연산 동작의 목표 입력 데이터를 사용할 필요가 없는 경우, 상기 목표 연산 동작이 완료된 후, 상기 목표 연산 동작에 대응하는 목표 입력 데이터의 저장 주소의 일부 또는 전부를 상기 목표 연산 동작의 목표 출력 데이터에 할당하도록 구성된다. 기타 실시예에서, 도 12에 제시된 바와 같이, 획득 모듈(410)은 메인 메모리의 가용 저장 용량, 슬레이브 메모리의 가용 저장 용량 및 목표 연산 동작을 획득하도록 구성되며; 입력 확정 모듈(420)은 상기 메인 메모리의 가용 저장 용량, 상기 슬레이브 메모리의 가용 저장 용량 및 상기 목표 연산 동작에 따라, 상기 목표 연산 동작에 대응하는 목표 입력 데이터를 확정하도록 구성되며; 출력 확정 모듈(430)은 상기 목표 연산 동작 및 상기 목표 입력 데이터에 따라, 상기 목표 연산 동작의 목표 출력 데이터를 확정하도록 구성되며; 저장 할당 모듈(440)은 상기 목표 연산 동작의 목표 출력 데이터가 상기 목표 연산 동작 후의 기타 연산 동작의 입력 데이터인 경우, 상기 목표 출력 데이터를 메인 메모리에 대응하여 저장하도록 구성된다. 선택적으로, 도 14에 제시된 바와 같이, 상기 데이터 전처리 장치는 저장 용량 확정 모듈(450)을 추가로 포함하며, 이는 상기 메인 메모리의 가용 저장 용량과 각 상기 슬레이브 메모리의 가용 저장 용량을 비교하여 최소의 가용 저장 용량을 제 1 메모리의 가용 저장 용량으로 사용하도록 구성되며; 입력 확정 모듈(420)은 상기 제 1 메모리의 가용 저장 용량 및 목표 연산 동작에 따라 목표 연산 동작에 대

응하는 목표 입력 데이터를 확정하도록 구성된다. 선택적으로, 상기 목표 연산 동작은 하나 이상의 연산 동작을 포함하고, 상기 연산 동작 각각은 그에 대응하는 서브 목표 입력 데이터가 있으며; 상기 입력 확정 모듈(420)은 융합 확정 유닛(421) 및 입력 확정 유닛(422)을 포함한다. 여기서 융합 확정 유닛(421)은 상기 제 1 메모리의 가용 저장 용량 및 상기 처리하고자 하는 연산에서 각 연산 동작의 융합 속성에 따라, 융합 가능한 연산 동작의 수량을 확정하여 융합 수량 역치를 얻도록 구성된다. 입력 확정 유닛(422)은 선정된 수량의 상기 융합 가능한 연산 동작의 조합을 상기 목표 연산 동작으로 사용하고, 상기 선정된 수량은 상기 융합 수량 역치이하 이며, 상기 선정된 수량의 융합 가능한 연산 동작 각각에 대응하는 서브 목표 입력 데이터를 상기 목표 연산 동작에 대응하는 목표 입력 데이터로 사용하도록 구성된다.

[0139] 선택적으로, 상기 처리하고자 하는 연산은 복수의 연산 계층을 포함하는 신경망 연산이며, 상기 연산 계층 각각은 하나의 상기 연산 동작을 나타내며; 상기 융합 확정 유닛(421)은 상기 신경망 연산의 각 연산 계층의 연결 관계에 따라 상기 각 연산 동작의 융합 속성을 확정하도록 구성된다. 선택적으로, 상기 목표 연산 동작은 하나 이상의 서브 목표 연산 동작을 포함하고, 상기 서브 목표 연산 동작 각각은 하나의 상기 목표 입력 데이터에 대응하며; 여기서 상기 목표 연산 동작에 대응하는 모든 입력 데이터는 복수의 입력 데이터 블록을 포함하고, 상기 목표 연산 동작에 대응하는 목표 입력 데이터의 수량은 하나 이상이며, 상기 목표 입력 데이터 각각은 하나 이상의 상기 입력 데이터 블록을 포함한다. 상기 입력 확정 모듈은 상기 서브 목표 연산 동작 각각의 목표 입력 데이터의 데이터 용량 및 목표 출력 데이터의 데이터 용량에 따라, 상기 서브 목표 연산 동작 각각에 필요한 목표 저장 용량을 확정하고; 상기 제 1 메모리의 가용 저장 용량 및 현재의 서브 목표 연산 동작에 필요한 목표 저장 용량에 따라 상기 제 1 메모리의 잔여 저장 용량을 확정하고; 상기 제 1 메모리의 잔여 저장 용량 및 상기 현재의 서브 목표 연산 동작 이외의 기타 서브 목표 연산 동작에 필요한 목표 저장 용량에 따라, 상기 서브 목표 연산 동작의 수량을 확정하도록 구성된다. 선택적으로, 상기 목표 입력 데이터는 제 1 목표 입력 데이터 및 제 2 목표 입력 데이터를 포함하고; 상기 입력 확정 모듈(420)은 기 설정된 연산 분배 규칙에 따라, 상기 메인 메모리에 대응하는 제 1 목표 입력 데이터와 상기 각 슬레이브 메모리에 대응하는 제 2 목표 입력 데이터를 확정하도록 구성되며; 상기 저장 할당 모듈(440)은 상기 메인 메모리의 가용 저장 용량 및 상기 제 1 목표 입력 데이터의 데이터 용량에 따라, 상기 메인 메모리 상에 있는 상기 제 1 목표 입력 데이터의 저장 주소를 확정하고; 상기 각 슬레이브 메모리의 가용 저장 용량 및 대응하는 상기 제 2 목표 입력 데이터의 데이터 용량에 따라, 상기 슬레이브 메모리 상에 있는 상기 제 2 목표 입력 데이터 각각의 저장 주소를 확정하도록 구성된다. 선택적으로, 상기 목표 출력 데이터는 제 1 목표 출력 데이터 및 제 2 목표 출력 데이터를 포함하고; 상기 출력 확정 모듈(430)은 상기 목표 연산 동작 및 상기 제 1 목표 입력 데이터에 따라, 상기 제 1 목표 출력 데이터 및 상기 제 1 목표 출력 데이터가 상기 메인 메모리 상에서 저장된 주소를 확정하도록 구성되고; 상기 목표 연산 동작 및 상기 제 2 목표 입력 데이터 각각에 따라,상기 제 2 목표 출력 데이터 및 상기 제 2 목표 출력 데이터 각각은 해당 슬레이브 메모리 상에서 저장된 주소를 각각 확정하도록 구성되고; 상기 제 2 목표 입력 데이터에 따라, 상기 제 2 목표 출력 데이터 각각은 상기 메인 메모리 상에 있는 저장된 주소를 확정하도록 구성된다. 선택적으로, 저장 할당 모듈(440)은 상기 슬레이브 프로세싱 회로에서 수행되는 기타 목표 연산 동작이 상기 제 2 목표 출력 데이터를 사용할 필요가 있을 때, 상기 제 2 목표 출력 데이터를 상기 슬레이브 프로세싱 회로에 대응하는 슬레이브 메모리에 저장하도록 구성된다. 또한, 상기 저장 할당 모듈(440)은 상기 목표 연산 동작의 목표 출력 데이터가 상기 목표 연산 동작 후에 기타 연산 동작의 입력 데이터인 경우, 상기 목표 출력 데이터를 상기 메인 메모리와 제 2 메모리에 대응하여 저장하도록 구성된다.

[0140] 상기 장치의 작동 원리는 상기 방법 중 각 단계의 실행 과정과 일치하다는 것을 분명히 해야 한다. 자세한 내용은 위의 설명을 참조할 수 있고, 상세한 설명은 여기서 생략한다.

[0141] 일 실시예에서, 본 개시의 실시예는 컴퓨터 프로그램을 저장한 컴퓨터 판독 가능 저장 매체도 제공하며, 상기 컴퓨터 프로그램이 프로세서에 의해 실행될 때 상기 실시예 중 임의 하나에 따른 방법의 단계를 구현한다. 구체적으로, 상기 컴퓨터 프로그램이 프로세서에 의해 실행될 때:

[0142] 제 1 메모리의 가용 저장 용량 및 목표 연산 동작을 획득하는 단계;

[0143] 목표 연산 동작 및 제 1 메모리의 가용 저장 용량에 따라, 상기 목표 연산 동작에 대응하는 목표 입력 데이터를 확정하고; 여기서 상기 목표 입력 데이터는 상기 목표 연산 동작에 대응하는 모든 입력 데이터의 일부 또는 전부인 단계;

[0144] 상기 목표 연산 동작 및 상기 목표 입력 데이터에 따라, 상기 목표 연산 동작의 목표 출력 데이터를 확정하는 단계;

[0145] 상기 목표 연산 동작의 목표 출력 데이터가 상기 목표 연산 동작 후의 기타 연산 동작의 입력 데이터인 경우, 상기 목표 연산 동작의 목표 출력 데이터를 상기 제 1 메모리 상에 저장하고, 여기서 상기 제 1 메모리는 프로세서에 인접하여 배치되는 단계를 구현한다.

[0146] 또한, 상기 프로세서는 메인 프로세싱 회로와 슬레이브 프로세싱 회로가 형성하는 마스터-슬레이브 구조일 수 있으며 이 경우, 상기 프로세서가 전술한 컴퓨터 프로그램을 실행할 때, :

[0147] 메인 메모리의 가용 저장 용량, 슬레이브 메모리의 가용 저장 용량 및 목표 연산 동작을 획득하는 단계;

[0148] 상기 메인 메모리의 가용 저장 용량, 상기 슬레이브 메모리의 가용 저장 용량 및 상기 목표 연산 동작에 따라, 상기 목표 연산 동작에 대응하는 목표 입력 데이터를 확정한다; 상기 목표 입력 데이터는 상기 목표 연산 동작에 대응하는 모든 입력 데이터의 일부 또는 전부인 단계;

[0149] 상기 목표 연산 동작 및 상기 목표 입력 데이터에 따라, 상기 목표 연산 동작에 대응하는 목표 출력 데이터를 확정하는 단계;

[0150] 상기 목표 연산 동작의 목표 출력 데이터가 상기 목표 연산 동작 후의 기타 연산 동작의 입력 데이터인 경우, 상기 목표 출력 데이터는 상기 메인 메모리에 대응하여 저장하는 단계를 구현한다.

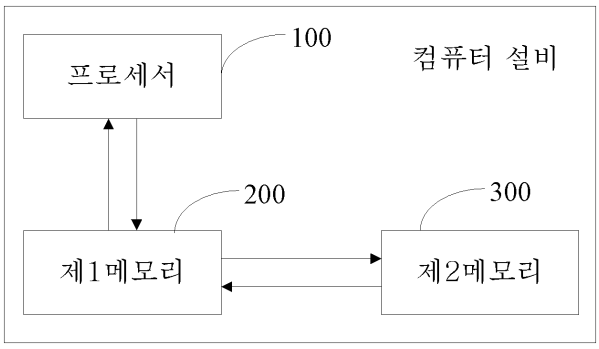
[0151] 상기 프로세서가 컴퓨터 프로그램을 실행하는 과정은 전술한 방법 중 각 단계의 실행 과정과 일치하다는 것을 알아야 한다. 자세한 내용은 위의 설명을 참조할 수 있고, 상세한 설명은 여기서 생략한다.

[0152] 상기 실시예의 각 기술적인 특징은 임의로 결합될 수 있으며, 설명의 간결함을 위해 상기 실시예의 각 기술적인 특징에 대한 모든 가능한 조합을 설명하지 않았지만, 이러한 기술적 특징의 조합에 모순이 존재하지 않는 한, 모두 본 명세서의 범위로 간주되어야 한다.

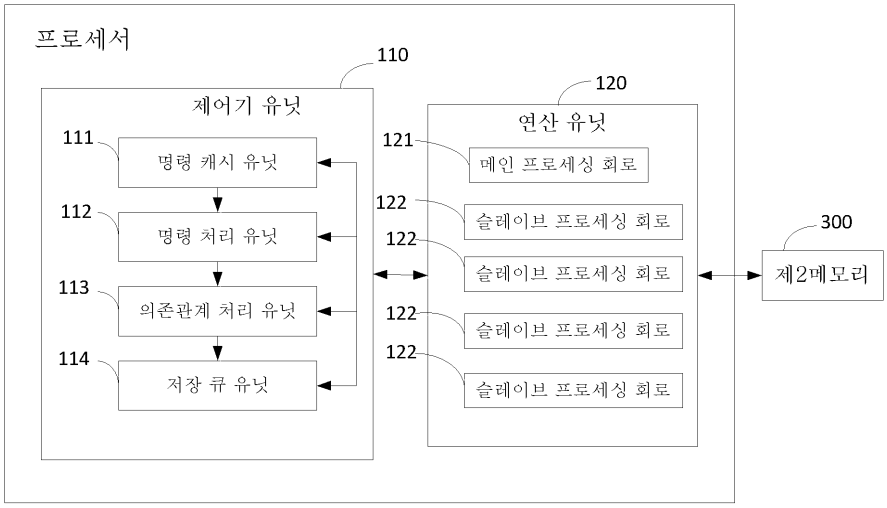
[0153] 상기 실시예는 본 개시의 몇몇 실시예를 설명한 것일 뿐이며, 그 설명은 비교적 구체적이고 상세하지만, 본 발명의 범위를 제한하는 것으로 해석되어서는 안된다. 당업자는, 본 발명의 사상을 벗어나지 않으면서 이루어진 다양한 변형 및 수정이 본 발명의 청구범위에 속한다. 그러므로 본 발명의 범위는 첨부된 청구 범위에 의해 결정된다.

도면

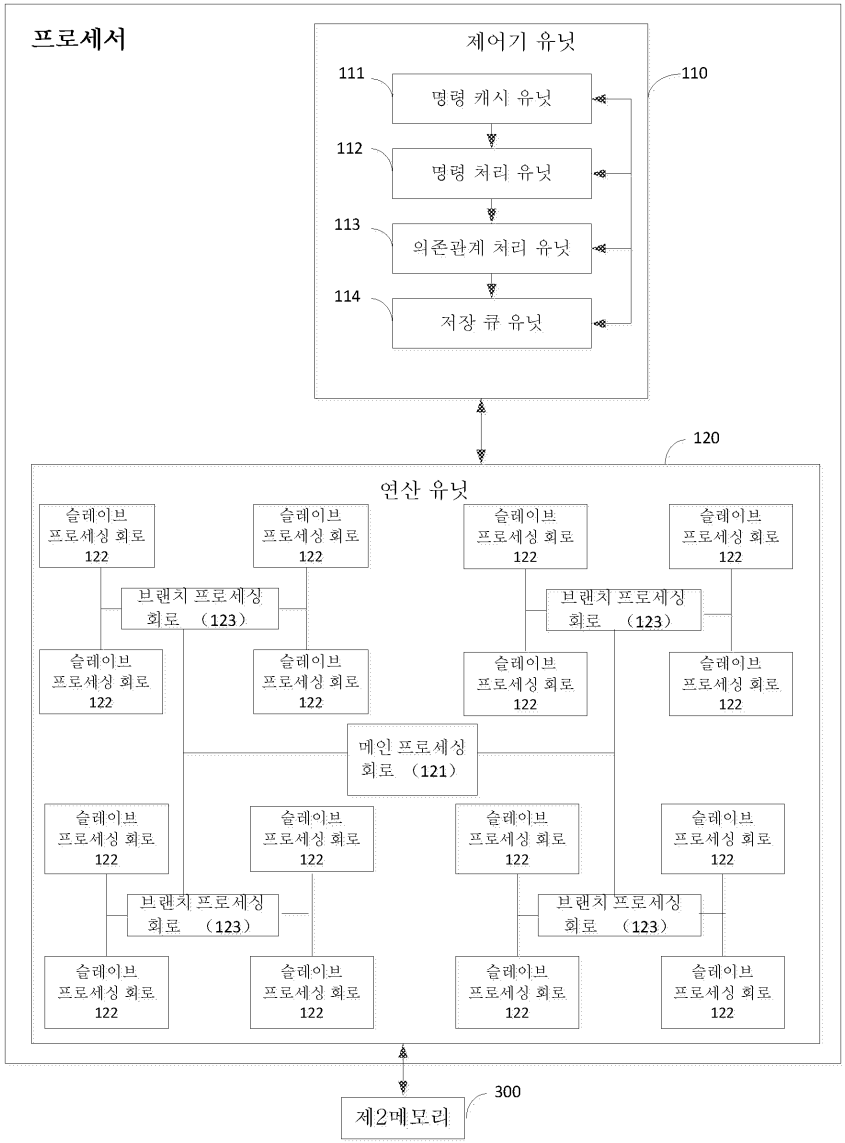
도면1



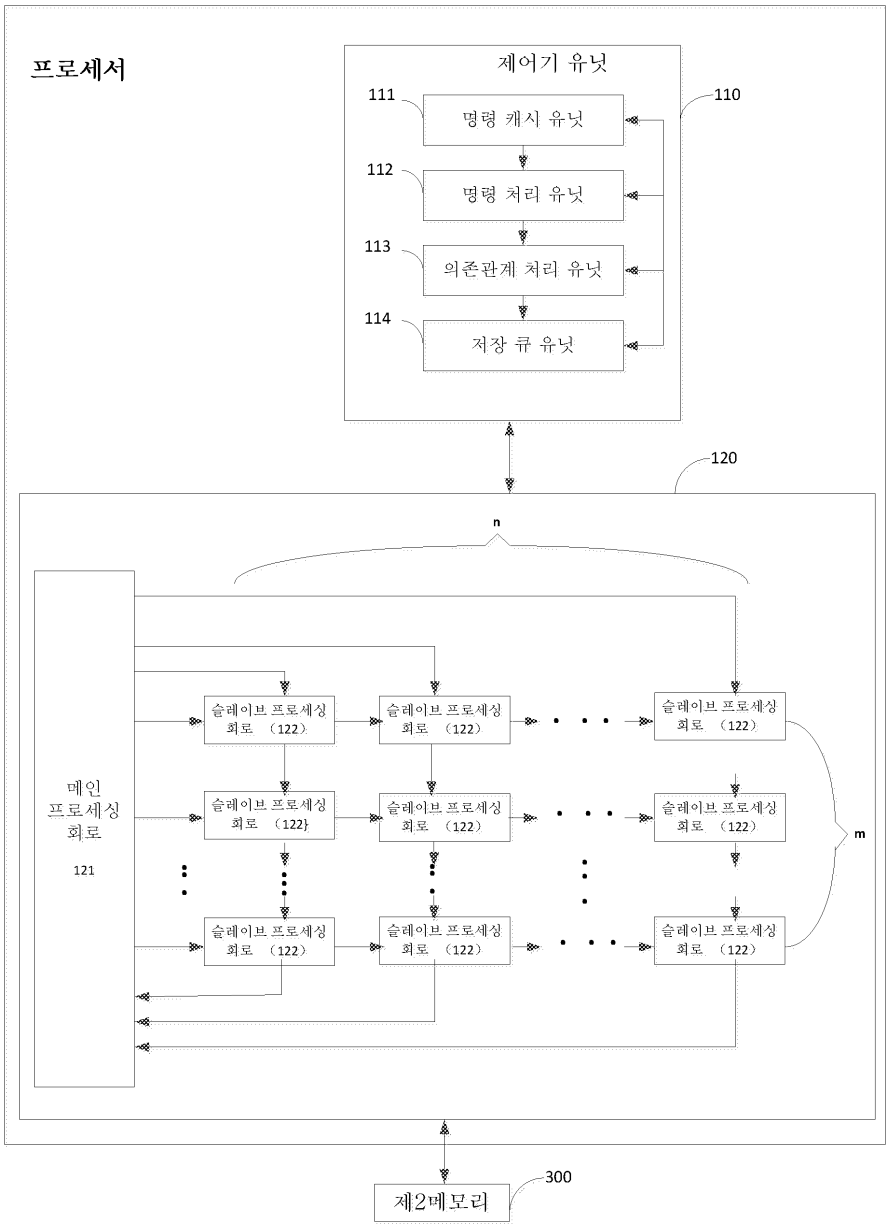
도면2


프로세서
110
제어기 유닛
111
명령 캐시 유닛
112
명령 처리 유닛
113
의존관계 처리 유닛
114
저장 큐 유닛
120
연산 유닛
121
메인 프로세싱 회로
122
슬레이브 프로세싱 회로
122
슬레이브 프로세싱 회로
122
슬레이브 프로세싱 회로
122
슬레이브 프로세싱 회로
300
제2메모리

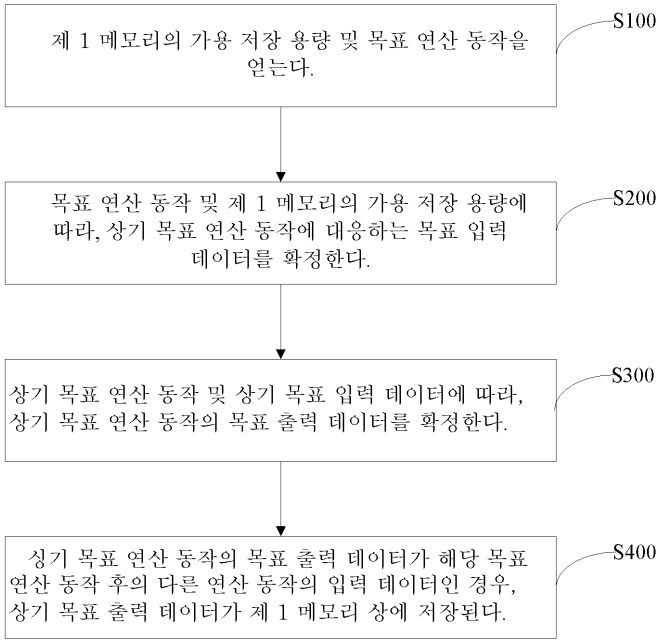
도면3


프로세서
제어기 유닛
111
명령 캐시 유닛
110
112
명령 처리 유닛
113
의존관계 처리 유닛
114
저장 큐 유닛
120
연산 유닛
슬레이브 프로세싱 회로 122
브랜치 프로세싱 회로 (123)
메인 프로세싱 회로 (121)
300
제2메모리

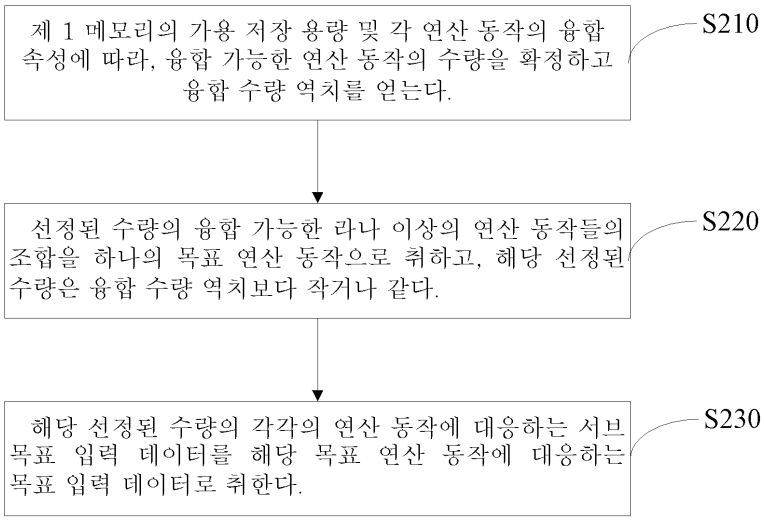
도면4


프로세서
제어기 유닛
111
명령 캐시 유닛
110
112
명령 처리 유닛
113
의존관계 처리 유닛
114
저장 큐 유닛
120
n
메인
프로세싱
회로
121
슬레이브 프로세싱
회로 (122)
m
제2메모리
300

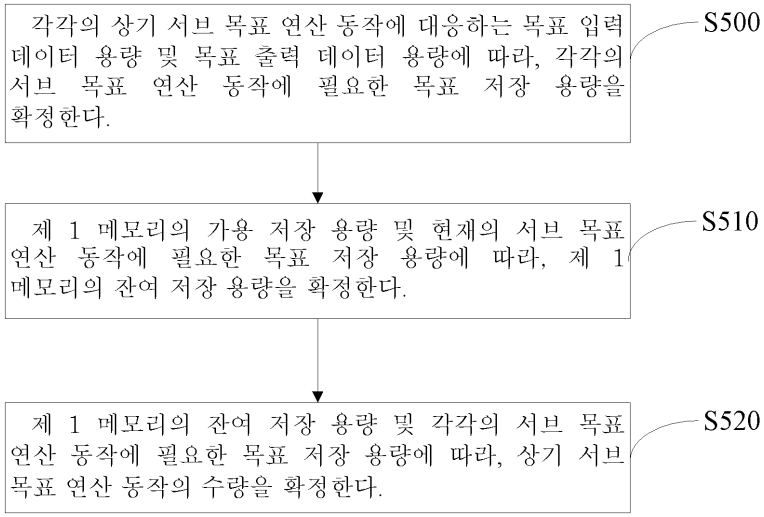
도면5


제 1 메모리의 가용 저장 용량 및 목표 연산 동작을 얻는다.
S100
목표 연산 동작 및 제 1 메모리의 가용 저장 용량에 따라, 상기 목표 연산 동작에 대응하는 목표 입력 데이터를 확정한다.
S200
상기 목표 연산 동작 및 상기 목표 입력 데이터에 따라, 상기 목표 연산 동작의 목표 출력 데이터를 확정한다.
S300
상기 목표 연산 동작의 목표 출력 데이터가 해당 목표 연산 동작 후의 다른 연산 동작의 입력 데이터인 경우, 상기 목표 출력 데이터가 제 1 메모리 상에 저장된다.
S400

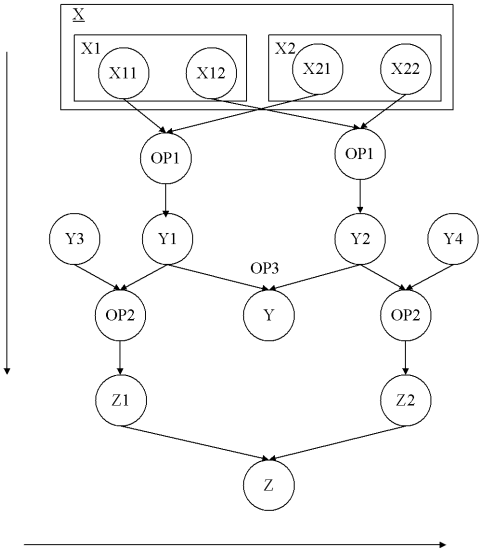
도면6


제 1 메모리의 가용 저장 용량 및 각 연산 동작의 융합 속성에 따라, 융합 가능한 연산 동작의 수량을 확정하고 융합 수량 역치를 얻는다.
S210
선정된 수량의 융합 가능한 라나 이상의 연산 동작들의 조합을 하나의 목표 연산 동작으로 취하고, 해당 선정된 수량은 융합 수량 역치보다 작거나 같다.
S220
해당 선정된 수량의 각각의 연산 동작에 대응하는 서브 목표 입력 데이터를 해당 목표 연산 동작에 대응하는 목표 입력 데이터로 취한다.
S230

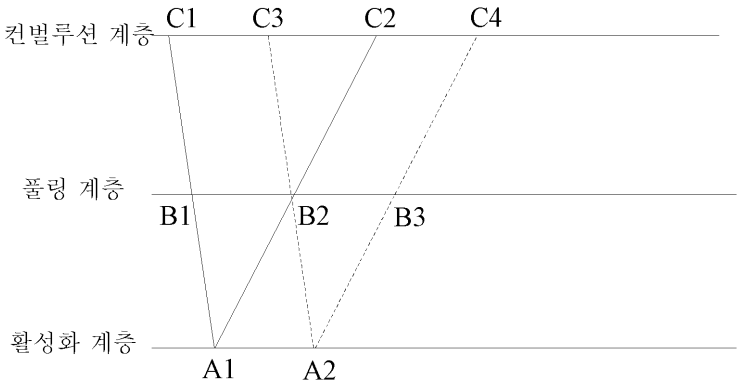
도면7


각각의 상기 서브 목표 연산 동작에 대응하는 목표 입력 데이터 용량 및 목표 출력 데이터 용량에 따라, 각각의 서브 목표 연산 동작에 필요한 목표 저장 용량을 확정한다.
S500
제 1 메모리의 가용 저장 용량 및 현재의 서브 목표 연산 동작에 필요한 목표 저장 용량에 따라, 제 1 메모리의 잔여 저장 용량을 확정한다.
S510
제 1 메모리의 잔여 저장 용량 및 각각의 서브 목표 연산 동작에 필요한 목표 저장 용량에 따라, 상기 서브 목표 연산 동작의 수량을 확정한다.
S520

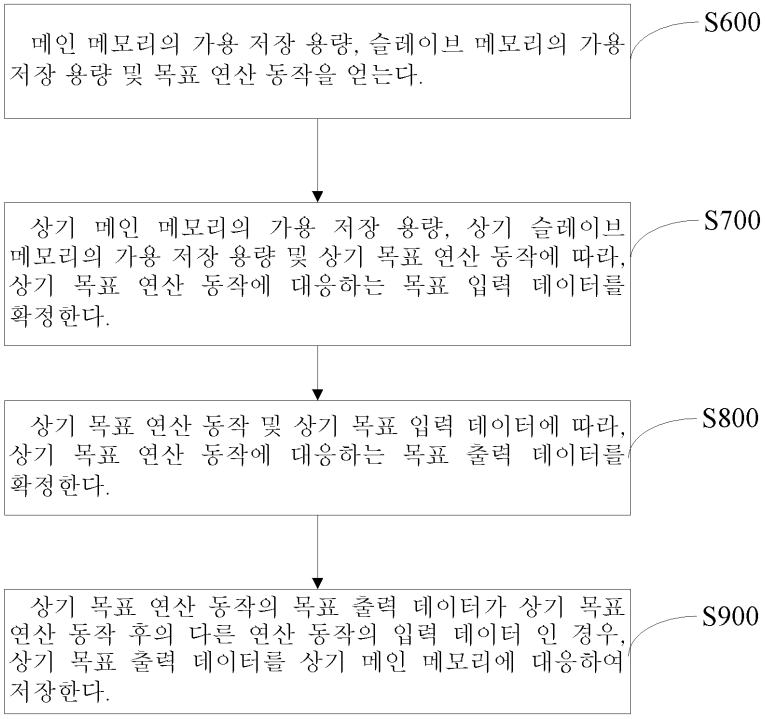
도면8


X
X1
X11
X12
X2
X21
X22
OP1
OP1
Y3
Y1
Y2
Y4
OP3
OP2
Y
OP2
Z1
Z2
Z

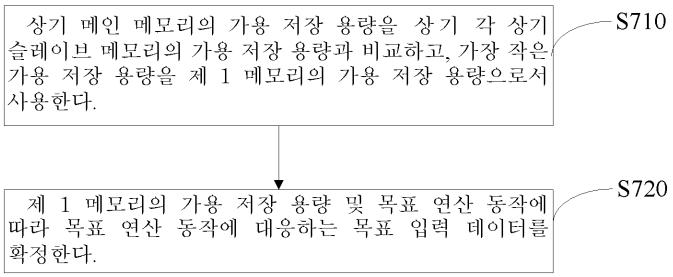
도면9


컨벌루션 계층
C1
C3
C2
C4
풀링 계층
B1
B2
B3
활성화 계층
A1
A2

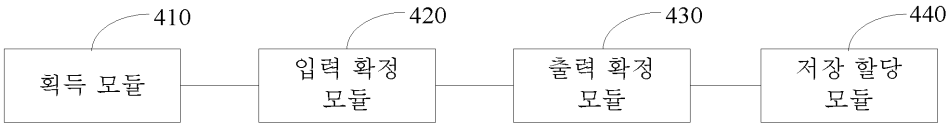
도면10



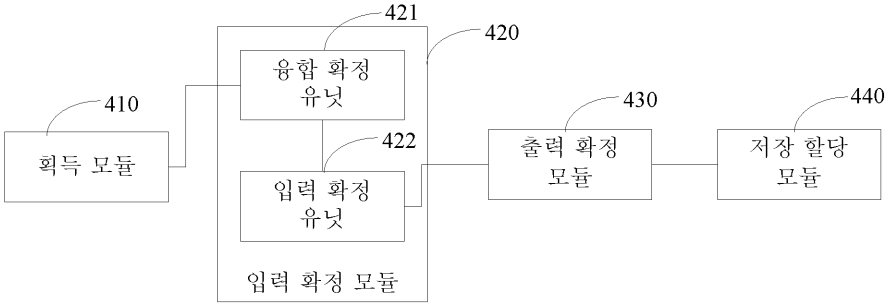
도면11



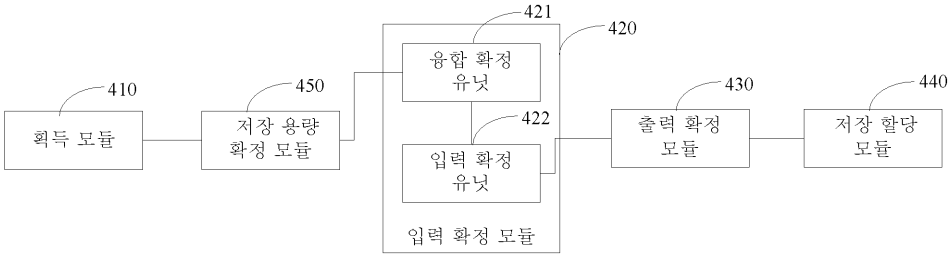
도면12



도면13



도면14


410
획득 모듈
450
저장 용량
확정 모듈
421
420
융합 확정
유닛
422
입력 확정
유닛
입력 확정 모듈
430
출력 확정
모듈
440
저장 할당
모듈