(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0103917
(43) 공개일자 2024년07월04일

(51) 국제특허분류(Int. Cl.)
G06N 3/096 (2023.01) *G06N 3/045* (2023.01)
(52) CPC특허분류
G06N 3/096 (2023.01)
G06N 3/045 (2023.01)
(21) 출원번호 10-2023-0033087
(22) 출원일자 2023년03월14일
심사청구일자 2023년03월14일
(30) 우선권주장
1020220184978 2022년12월26일 대한민국(KR)

(71) 출원인
경희대학교 산학협력단
경기도 용인시 기흥구 덕영대로 1732 (서천동, 경희대학교 국제캠퍼스내)
(72) 발명자
허의남
경기도 용인시 기흥구 동탄기흥로 864, 106동 401호(공세동, 공세 호수마을 공세성원상떼레이크뷰)
샤르만 아카타르
쿠슈티아, 다울랏푸르, 라파예트퍼, 호링가치
서유리
서울특별시 영등포구 가마산로46길 8, 402호(대림동, SKE캐슬)
(74) 대리인
두호특허법인

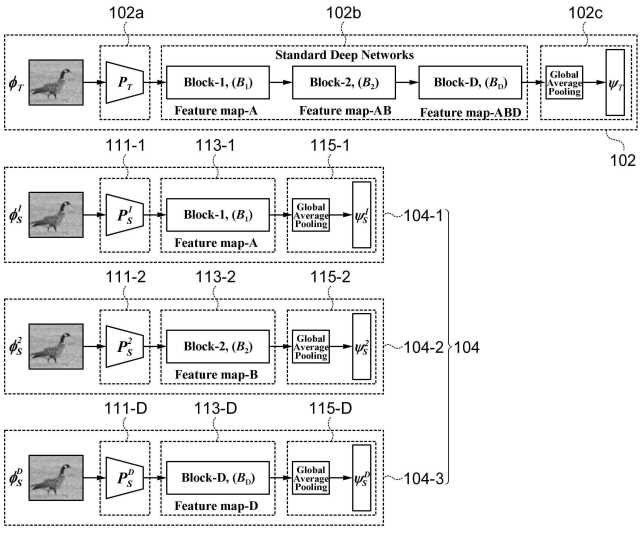
전체 청구항 수 : 총 19 항

(54) 발명의 명칭 **지식 증류 기반의 인공 신경망 학습 방법 및 이를 수행하기 위한 컴퓨팅 장치**

(57) 요 약

지식 증류 기반의 인공 신경망 학습 방법 및 이를 수행하기 위한 컴퓨팅 장치가 개시된다. 개시되는 일 실시예에 따른 인공 신경망 학습 방법은, 입력되는 대상 이미지를 일정 크기로 전처리 하는 전처리부, 전처리 된 대상 이미지로부터 특징 맵을 추출하는 복수 개의 순차적인 블록을 구비하는 특징 추출부, 및 특징 맵에 기초하여 예측값을 출력하는 출력부를 포함하는 선생 모델을 학습시키는 단계, 선생 모델의 복수 개의 순차적인 블록을 별개의 블록으로 각각 분할하는 단계, 분할된 복수 개의 블록들을 병렬로 배치하여 복수 개의 서브 학생 모델을 생성하는 단계, 및 복수 개의 서브 학생 모델을 각각 학습시키는 단계를 포함한다.

대 표 도



이 발명을 지원한 국가연구개발사업

과제고유번호	1711160469
과제번호	2022-0-00047-001
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	SW컴퓨팅산업원천기술개발
연구과제명	응용서비스 운영 지능화를 지원하는 마이크로서비스개발·운영 플랫폼 기술 개발
기 여 율	1/1
과제수행기관명	오케스트로 주식회사
연구기간	2022.04.01 ~ 2022.12.31

명 세 서

청구범위

청구항 1

하나 이상의 프로세서들, 및

상기 하나 이상의 프로세서들에 의해 실행되는 하나 이상의 프로그램들을 저장하는 메모리를 구비한 컴퓨팅 장치에서 수행되는 방법으로,

입력되는 대상 이미지를 일정 크기로 전처리 하는 전처리부, 전처리 된 대상 이미지로부터 특징 맵을 추출하는 복수 개의 순차적인 블록을 구비하는 특징 추출부, 및 특징 맵에 기초하여 예측 값을 출력하는 출력부를 포함하는 선생 모델을 학습시키는 단계;

상기 선생 모델의 복수 개의 순차적인 블록을 별개의 블록으로 각각 분할하는 단계;

상기 분할된 복수 개의 블록들을 병렬로 배치하여 복수 개의 서브 학생 모델을 생성하는 단계; 및

상기 복수 개의 서브 학생 모델을 각각 학습시키는 단계를 포함하는, 인공 신경망 학습 방법.

청구항 2

청구항 1에 있어서,

상기 복수 개의 서브 학생 모델을 생성하는 단계는,

상기 분할된 복수 개의 블록들을 상기 병렬로 배치된 복수 개의 서브 학생 모델의 각 특징 추출부에 순서에 따라 배치하는 단계를 포함하는, 인공 신경망 학습 방법.

청구항 3

청구항 2에 있어서,

상기 복수 개의 서브 학생 모델은,

각각 동일한 대상 이미지를 입력으로 하는, 인공 신경망 학습 방법.

청구항 4

청구항 3에 있어서,

상기 복수 개의 서브 학생 모델을 생성하는 단계는,

각 서브 학생 모델에서 각 특징 추출부의 전단에 대상 이미지를 서로 다른 크기로 전처리 하는 전처리부를 배치하는 단계를 더 포함하는, 인공 신경망 학습 방법.

청구항 5

청구항 4에 있어서,

상기 전처리부를 배치하는 단계는,

첫 번째 서브 학생 모델은 상기 선생 모델의 전처리부와 동일한 크기로 대상 이미지를 전처리 할 수 있는 전처리부를 배치하는 단계; 및

n(n은 2 이상 이고 상기 선생 모델의 블록의 전체 개수 이하인 자연수) 번째 서브 학생 모델은 상기 선생 모델의 n-1 번째 블록에서 출력하는 특징맵의 크기와 동일한 크기로 대상 이미지를 전처리 할 수 있는 전처리부를 배치하는 단계를 포함하는, 인공 신경망 학습 방법.

청구항 6

청구항 5에 있어서,

상기 복수 개의 서브 학생 모델을 생성하는 단계는,

각 서브 학생 모델의 각 특징 추출부의 후단에 해당 특징 추출부로부터 출력되는 특징 맵에 기초하여 상기 대상 이미지에 대한 예측 값을 출력하는 출력부를 배치하는 단계를 더 포함하는, 인공 신경망 학습 방법.

청구항 7

청구항 6에 있어서,

상기 복수 개의 서브 학생 모델은,

각 서브 학생 모델의 출력부에서 출력하는 예측 값과 기 설정된 정답 값 간의 손실 함수인 제1 손실 함수, 각 서브 학생 모델의 출력부에서 출력하는 예측 값과 상기 선생 모델의 출력부에서 출력하는 예측 값 간의 손실 함수인 제2 손실 함수, 및 각 서브 학생 모델의 특징 추출부의 블록에서 생성한 특징 맵과 상기 선생 모델의 그에 대응하는 특징 추출부의 블록에서 생성한 특징 맵 간의 손실 함수인 제3 손실 함수를 가지는, 인공 신경망 학습 방법.

청구항 8

청구항 7에 있어서,

상기 각 서브 학생 모델의 전체 손실 함수는 하기 수학식으로 나타내는, 인공 신경망 학습 방법.

(수학식)

$$\mathcal{L} = \lambda_1 \mathcal{L}_{ce} + \lambda_2 \mathcal{L}_{kd} + \mathcal{L}_{feature}$$

$\mathcal{L}_{ce}$: 제1 손실 함수

$\mathcal{L}_{kd}$: 제2 손실 함수

$\mathcal{L}_{feature}$: 제3 손실 함수

λ_1, λ_2 : 기 설정된 가중치

청구항 9

청구항 1에 있어서,

상기 인공 신경망 학습 방법은,

상기 복수 개의 서브 학생 모델의 학습이 완료된 경우, 대상 이미지가 상기 복수 개의 서브 학생 모델로 각각 입력되면, 상기 복수 개의 서브 학생 모델에서 출력하는 예측 값들에 기초하여 최종 예측 값을 산출하는 단계를 더 포함하는, 인공 신경망 학습 방법.

청구항 10

하나 이상의 프로세서들;

메모리; 및

하나 이상의 프로그램들을 포함하고,

상기 하나 이상의 프로그램들은 상기 메모리에 저장되고, 상기 하나 이상의 프로세서들에 의해 실행되도록 구성되며,

상기 하나 이상의 프로그램들은,

입력되는 대상 이미지를 일정 크기로 전처리 하는 전처리부, 전처리 된 대상 이미지로부터 특징 맵을 추출하는 복수 개의 순차적인 블록을 구비하는 특징 추출부, 및 특징 맵에 기초하여 예측 값을 출력하는 출력부를 포함하는 선생 모델을 학습시키기 위한 명령;

상기 선생 모델의 복수 개의 순차적인 블록을 별개의 블록으로 각각 분할하기 위한 명령;

상기 분할된 복수 개의 블록들을 병렬로 배치하여 복수 개의 서브 학생 모델을 생성하기 위한 명령; 및

상기 복수 개의 서브 학생 모델을 각각 학습시키기 위한 명령을 포함하는, 컴퓨팅 장치.

청구항 11

청구항 10에 있어서,

상기 복수 개의 서브 학생 모델을 생성하기 위한 명령은,

상기 분할된 복수 개의 블록들을 상기 병렬로 배치된 복수 개의 서브 학생 모델의 각 특징 추출부에 순서에 따라 배치하기 위한 명령을 포함하는, 컴퓨팅 장치.

청구항 12

청구항 11에 있어서,

상기 복수 개의 서브 학생 모델은,

각각 동일한 대상 이미지를 입력으로 하는, 컴퓨팅 장치.

청구항 13

청구항 12에 있어서,

상기 복수 개의 서브 학생 모델을 생성하기 위한 명령은,

각 서브 학생 모델에서 각 특징 추출부의 전단에 대상 이미지를 서로 다른 크기로 전처리 하는 전처리부를 배치하기 위한 명령을 더 포함하는, 컴퓨팅 장치.

청구항 14

청구항 13에 있어서,

상기 전처리부를 배치하기 위한 명령은,

첫 번째 서브 학생 모델은 상기 선생 모델의 전처리부와 동일한 크기로 대상 이미지를 전처리 할 수 있는 전처

리부를 배치하기 위한 명령; 및

n(n은 2 이상 이고 상기 선생 모델의 블록의 전체 개수 이하인 자연수) 번째 서브 학생 모델은 상기 선생 모델의 n-1 번째 블록에서 출력하는 특징맵의 크기와 동일한 크기로 대상 이미지를 전처리 할 수 있는 전처리부를 배치하기 위한 명령을 포함하는, 컴퓨팅 장치.

청구항 15

청구항 14에 있어서,

상기 복수 개의 서브 학생 모델을 생성하기 위한 명령은,

각 서브 학생 모델의 각 특징 추출부의 후단에 해당 특징 추출부로부터 출력되는 특징 맵에 기초하여 상기 대상 이미지에 대한 예측 값을 출력하는 출력부를 배치하기 위한 명령을 더 포함하는, 컴퓨팅 장치.

청구항 16

청구항 15에 있어서,

상기 복수 개의 서브 학생 모델은,

각 서브 학생 모델의 출력부에서 출력하는 예측 값과 기 설정된 정답 값 간의 손실 함수인 제1 손실 함수, 각 서브 학생 모델의 출력부에서 출력하는 예측 값과 상기 선생 모델의 출력부에서 출력하는 예측 값 간의 손실 함수인 제2 손실 함수, 및 각 서브 학생 모델의 특징 추출부의 블록에서 생성한 특징 맵과 상기 선생 모델의 그에 대응하는 특징 추출부의 블록에서 생성한 특징 맵 간의 손실 함수인 제3 손실 함수를 가지는, 컴퓨팅 장치.

청구항 17

청구항 16에 있어서,

상기 각 서브 학생 모델의 전체 손실 함수는 하기 수학식으로 나타내는, 컴퓨팅 장치.

(수학식)

$$\mathcal{L} = \lambda_1 \mathcal{L}_{ce} + \lambda_2 \mathcal{L}_{kd} + \mathcal{L}_{feature}$$

$\mathcal{L}_{ce}$: 제1 손실 함수

$\mathcal{L}_{kd}$: 제2 손실 함수

$\mathcal{L}_{feature}$: 제3 손실 함수

λ_1, λ_2 : 기 설정된 가중치

청구항 18

청구항 10에 있어서,

상기 하나 이상의 프로그램들은,

상기 복수 개의 서브 학생 모델의 학습이 완료된 경우, 대상 이미지가 상기 복수 개의 서브 학생 모델로 각각 입력되면, 상기 복수 개의 서브 학생 모델에서 출력하는 예측 값들에 기초하여 최종 예측 값을 산출하기 위한 명령을 더 포함하는, 컴퓨팅 장치.

청구항 19

비일시적 컴퓨터 판독 가능한 저장 매체(non-transitory computer readable storage medium)에 저장된 컴퓨터 프로그램으로서,

상기 컴퓨터 프로그램은 하나 이상의 명령어들을 포함하고, 상기 명령어들은 하나 이상의 프로세서들을 갖는 컴퓨팅 장치에 의해 실행될 때, 상기 컴퓨팅 장치로 하여금,

입력되는 대상 이미지를 일정 크기로 전처리 하는 전처리부, 전처리 된 대상 이미지로부터 특징 맵을 추출하는 복수 개의 순차적인 블록을 구비하는 특징 추출부, 및 특징 맵에 기초하여 예측 값을 출력하는 출력부를 포함하는 선생 모델을 학습시키는 단계;

상기 선생 모델의 복수 개의 순차적인 블록을 별개의 블록으로 각각 분할하는 단계;

상기 분할된 복수 개의 블록들을 병렬로 배치하여 복수 개의 서브 학생 모델을 생성하는 단계; 및

상기 복수 개의 서브 학생 모델을 각각 학습시키는 단계를 수행하도록 하는, 컴퓨터 프로그램.

발명의 설명

기 술 분 야

[0001] 본 발명의 실시예는 지식 증류 기반의 인공 신경망 아키텍처와 관련된다.

배 경 기 술

[0003] 최근, 딥 러닝(deep learning) 기반 기술은 감시 시스템, 자율 주행, 엣지 장치의 실시간 컴퓨팅과 같은 다양한 분야 및 작업에서 높은 성능을 보여주고 있다. 그러나, 고성능의 딥 러닝 모델은 수백만 개의 매개 변수를 가지고 있고 연산 량이 많기 때문에 많은 컴퓨팅 자원이 요구된다. 따라서, 컴퓨팅 자원에 한계가 있는 장치(예를 들어, 사물 인터넷 장치, 엣지 장치, 또는 모바일 장치 등)에서 딥 러닝 기반의 애플리케이션을 실시간 수행하는 것은 어려움이 있다.

선행기술문헌

특허문헌

[0005] (특허문헌 0001) 한국등록특허공보 제10-2232138호(2021.03.25)

발명의 내용

해결하려는 과제

[0006] 개시되는 실시예는 제한된 컴퓨팅 자원을 가진 장치에서 애플리케이션을 실시간으로 실행할 수 있도록 하는 지식 증류 기반의 인공 신경망 학습 방법 및 이를 수행하기 위한 컴퓨팅 장치를 제공하기 위한 것이다.

과제의 해결 수단

[0008] 개시되는 일 실시예에 따른 인공 신경망 학습 방법은, 하나 이상의 프로세서들, 및 상기 하나 이상의 프로세서들에 의해 실행되는 하나 이상의 프로그램들을 저장하는 메모리를 구비한 컴퓨팅 장치에서 수행되는 방법으로, 입력되는 대상 이미지를 일정 크기로 전처리 하는 전처리부, 전처리 된 대상 이미지로부터 특징 맵을 추출하는

복수 개의 순차적인 블록을 구비하는 특징 추출부, 및 특징 맵에 기초하여 예측 값을 출력하는 출력부를 포함하는 선생 모델을 학습시키는 단계; 상기 선생 모델의 복수 개의 순차적인 블록을 별개의 블록으로 각각 분할하는 단계; 상기 분할된 복수 개의 블록들을 병렬로 배치하여 복수 개의 서브 학생 모델을 생성하는 단계; 및 상기 복수 개의 서브 학생 모델을 각각 학습시키는 단계를 포함한다.

[0009] 상기 복수 개의 서브 학생 모델을 생성하는 단계는, 상기 분할된 복수 개의 블록들을 상기 병렬로 배치된 복수 개의 서브 학생 모델의 각 특징 추출부에 순서에 따라 배치하는 단계를 포함할 수 있다.

[0010] 상기 복수 개의 서브 학생 모델은, 각각 동일한 대상 이미지를 입력으로 할 수 있다.

[0011] 상기 복수 개의 서브 학생 모델을 생성하는 단계는, 각 서브 학생 모델에서 각 특징 추출부의 전단에 대상 이미지를 서로 다른 크기로 전처리 하는 전처리부를 배치하는 단계를 더 포함할 수 있다.

[0012] 상기 전처리부를 배치하는 단계는, 첫 번째 서브 학생 모델은 상기 선생 모델의 전처리부와 동일한 크기로 대상 이미지를 전처리 할 수 있는 전처리부를 배치하는 단계; 및 n(n은 2 이상 이고 상기 선생 모델의 블록의 전체 개수 이하인 자연수) 번째 서브 학생 모델은 상기 선생 모델의 n-1 번째 블록에서 출력하는 특징맵의 크기와 동일한 크기로 대상 이미지를 전처리 할 수 있는 전처리부를 배치하는 단계를 포함할 수 있다.

[0013] 상기 복수 개의 서브 학생 모델을 생성하는 단계는, 각 서브 학생 모델의 각 특징 추출부의 후단에 해당 특징 추출부로부터 출력되는 특징 맵에 기초하여 상기 대상 이미지에 대한 예측 값을 출력하는 출력부를 배치하는 단계를 더 포함할 수 있다.

[0014] 상기 복수 개의 서브 학생 모델은, 각 서브 학생 모델의 출력부에서 출력하는 예측 값과 기 설정된 정답 값 간의 손실 함수인 제1 손실 함수, 각 서브 학생 모델의 출력부에서 출력하는 예측 값과 상기 선생 모델의 출력부에서 출력하는 예측 값 간의 손실 함수인 제2 손실 함수, 및 각 서브 학생 모델의 특징 추출부의 블록에서 생성한 특징 맵과 상기 선생 모델의 그에 대응하는 특징 추출부의 블록에서 생성한 특징 맵 간의 손실 함수인 제3 손실 함수를 가질 수 있다.

[0015] 상기 각 서브 학생 모델의 전체 손실 함수는 하기 수학식으로 나타낼 수 있다.

[0016] (수학식)

[0017] $$\mathcal{L} = \lambda_1 \mathcal{L}_{ce} + \lambda_2 \mathcal{L}_{kd} + \mathcal{L}_{feature}$$

[0018] $\mathcal{L}_{ce}$: 제1 손실 함수

[0019] $\mathcal{L}_{kd}$: 제2 손실 함수

[0020] $\mathcal{L}_{feature}$: 제3 손실 함수

[0021] λ_1, λ_2 : 기 설정된 가중치

[0022] 상기 인공 신경망 학습 방법은, 상기 복수 개의 서브 학생 모델의 학습이 완료된 경우, 대상 이미지가 상기 복수 개의 서브 학생 모델로 각각 입력되면, 상기 복수 개의 서브 학생 모델에서 출력하는 예측 값들에 기초하여 최종 예측 값을 산출하는 단계를 더 포함할 수 있다.

[0023] 개시되는 일 실시예에 따른 컴퓨팅 장치는, 하나 이상의 프로세서들; 메모리; 및 하나 이상의 프로그램들을 포함하고, 상기 하나 이상의 프로그램들은 상기 메모리에 저장되고, 상기 하나 이상의 프로세서들에 의해 실행되도록 구성되며, 상기 하나 이상의 프로그램들은, 입력되는 대상 이미지를 일정 크기로 전처리 하는 전처리부, 전처리 된 대상 이미지로부터 특징 맵을 추출하는 복수 개의 순차적인 블록을 구비하는 특징 추출부, 및 특징 맵에 기초하여 예측 값을 출력하는 출력부를 포함하는 선생 모델을 학습시키기 위한 명령; 상기 선생 모델의 복수 개의 순차적인 블록을 별개의 블록으로 각각 분할하기 위한 명령; 상기 분할된 복수 개의 블록들을 병렬로 배치하여 복수 개의 서브 학생 모델을 생성하기 위한 명령; 및 상기 복수 개의 서브 학생 모델을 각각 학습시키기 위한 명령을 포함한다.

[0024] 상기 복수 개의 서브 학생 모델을 생성하기 위한 명령은, 상기 분할된 복수 개의 블록들을 상기 병렬로 배치된

복수 개의 서브 학생 모델의 각 특징 추출부에 순서에 따라 배치하기 위한 명령을 포함할 수 있다.

[0025] 상기 복수 개의 서브 학생 모델은, 각각 동일한 대상 이미지를 입력으로 할 수 있다.

[0026] 상기 복수 개의 서브 학생 모델을 생성하기 위한 명령은, 각 서브 학생 모델에서 각 특징 추출부의 전단에 대상 이미지를 서로 다른 크기로 전처리 하는 전처리부를 배치하기 위한 명령을 더 포함할 수 있다.

[0027] 상기 전처리부를 배치하기 위한 명령은, 첫 번째 서브 학생 모델은 상기 선생 모델의 전처리부와 동일한 크기로 대상 이미지를 전처리 할 수 있는 전처리부를 배치하기 위한 명령; 및 n(n은 2 이상 이고 상기 선생 모델의 블록의 전체 개수 이하인 자연수) 번째 서브 학생 모델은 상기 선생 모델의 n-1 번째 블록에서 출력하는 특징맵의 크기와 동일한 크기로 대상 이미지를 전처리 할 수 있는 전처리부를 배치하기 위한 명령을 포함할 수 있다.

[0028] 상기 복수 개의 서브 학생 모델을 생성하기 위한 명령은, 각 서브 학생 모델의 각 특징 추출부의 후단에 해당 특징 추출부로부터 출력되는 특징 맵에 기초하여 상기 대상 이미지에 대한 예측 값을 출력하는 출력부를 배치하기 위한 명령을 더 포함할 수 있다.

[0029] 상기 복수 개의 서브 학생 모델은, 각 서브 학생 모델의 출력부에서 출력하는 예측 값과 기 설정된 정답 값 간의 손실 함수인 제1 손실 함수, 각 서브 학생 모델의 출력부에서 출력하는 예측 값과 상기 선생 모델의 출력부에서 출력하는 예측 값 간의 손실 함수인 제2 손실 함수, 및 각 서브 학생 모델의 특징 추출부의 블록에서 생성한 특징 맵과 상기 선생 모델의 그에 대응하는 특징 추출부의 블록에서 생성한 특징 맵 간의 손실 함수인 제3 손실 함수를 가질 수 있다.

[0030] 상기 하나 이상의 프로그램들은, 상기 복수 개의 서브 학생 모델의 학습이 완료된 경우, 대상 이미지가 상기 복수 개의 서브 학생 모델로 각각 입력되면, 상기 복수 개의 서브 학생 모델에서 출력하는 예측 값들에 기초하여 최종 예측 값을 산출하기 위한 명령을 더 포함할 수 있다.

발명의 효과

[0032] 개시되는 실시예에 의하면, 복수 개의 서브 학생 모델이 병렬로 배치되기 때문에, 태스크에 대한 전체 실행 시간을 줄일 수 있으며, 그로 인해 실시간 서비스가 필요한 애플리케이션에 적용할 수 있게 된다.

도면의 간단한 설명

[0034] 도 1은 본 발명의 일 실시예에 따른 인공 신경망 아키텍처를 나타낸 도면

도 2는 본 발명의 일 실시예에서 제1 서브 학생 모델과 선생 모델 간의 제2 손실 함수 및 제3 손실 함수를 나타낸 도면

도 3은 본 발명의 일 실시예에서 선생 모델의 전체 실행 시간과 학생 모델의 전체 실행 시간을 비교한 도면

도 4는 예시적인 실시예들에서 사용되기에 적합한 컴퓨팅 장치를 포함하는 컴퓨팅 환경을 예시하여 설명하기 위한 블록도

도 5는 본 발명의 일 실시예에 따른 인공 신경망의 학습 방법을 나타낸 흐름도

발명을 실시하기 위한 구체적인 내용

[0035] 이하, 도면을 참조하여 본 발명의 구체적인 실시형태를 설명하기로 한다. 이하의 상세한 설명은 본 명세서에서 기술된 방법, 장치 및/또는 시스템에 대한 포괄적인 이해를 돕기 위해 제공된다. 그러나 이는 예시에 불과하며 본 발명은 이에 제한되지 않는다.

[0036] 본 발명의 실시예들을 설명함에 있어서, 본 발명과 관련된 공지기술에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략하기로 한다. 그리고, 후술되는 용어들은 본 발명에서의 기능을 고려하여 정의된 용어들로서 이는 사용자, 운용자의 의도 또는 관례 등에 따라 달라질 수 있다. 그러므로 그 정의는 본 명세서 전반에 걸친 내용을 토대로 내려져야 할 것이다. 상세한 설명에서 사용되는 용어는 단지 본 발명의 실시예들을 기술하기 위한 것이며, 결코 제한적이어서는 안 된다. 명확하게 달리 사용되지 않는 한, 단수 형태의 표현은 복수 형태의 의미를 포함한다. 본 설명에서, "포함" 또는 "구비"와 같은

표현은 어떤 특성들, 숫자들, 단계들, 동작들, 요소들, 이들의 일부 또는 조합을 가리키기 위한 것이며, 기술된 것 이외에 하나 또는 그 이상의 다른 특성, 숫자, 단계, 동작, 요소, 이들의 일부 또는 조합의 존재 또는 가능성을 배제하도록 해석되어서는 안 된다.

[0037] 또한, 제1, 제2 등의 용어는 다양한 구성 요소들을 설명하는데 사용될 수 있지만, 상기 구성 요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성 요소를 다른 구성 요소로부터 구별하는 목적으로 사용될 수 있다. 예를 들어, 본 발명의 권리 범위를 벗어나지 않으면서 제1 구성 요소는 제2 구성 요소로 명명될 수 있고, 유사하게 제2 구성 요소도 제1 구성 요소로 명명될 수 있다.

[0039] 도 1은 본 발명의 일 실시예에 따른 인공 신경망 아키텍처를 나타낸 도면이다.

[0040] 도 1을 참조하면, 인공 신경망 아키텍처(100)는 선생 모델(102) 및 학생 모델(104)을 포함할 수 있다. 인공 신경망 아키텍처(100)는 지식 증류(knowledge distillation) 기반의 인공 신경망 아키텍처 일 수 있다.

[0041] 선생 모델(102)은 학생 모델(104)이 수행하고자 하는 태스크에 대해 이미 학습된 기계 학습 모델로서 학생 모델(104)에 비하여 신경망의 크기가 큰 모델을 의미할 수 있다. 여기서, 신경망의 크기가 크다는 것은 신경망 레이어, 커널, 노드, 또는 파라미터 등의 개수가 보다 많다는 것을 의미할 수 있다. 학생 모델(104)은 기 학습된 선생 모델(102)에서 습득한 지식을 활용하여 학습되는 기계 학습 모델로서 선생 모델(102)에 비하여 신경망의 크기가 작은 모델을 의미할 수 있다.

[0042] 선생 모델(102)은 전처리부(102a), 특징 추출부(102b), 및 출력부(102c)를 포함할 수 있다. 전처리부(102a)는 대상 이미지를 입력 받아 전처리를 수행할 수 있다. 전처리부(102a)는 대상 이미지를 기 설정된 채널 수로 변환할 수 있다.

[0043] 특징 추출부(102b)는 대상 이미지로부터 특징(feature)을 추출하는 신경망일 수 있다. 특징 추출부(102b)의 신경망은 복수 개의 블록(Block-1 ~ Block-D)을 포함할 수 있다. D는 블록의 전체 개수이다. 복수 개의 블록(Block-1 ~ Block-D)은 순차적으로 연결될 수 있다. 즉, 제1 블록(Block-1)의 출력은 제2 블록(Block-2)으로 입력되고, 제2 블록(Block-2)의 출력은 제3 블록(Block-3)으로 입력될 수 있다. 여기서는, 설명의 편의상 블록의 개수가 3개인 것을 일 예로 도시하였다. 그리고, 각 블록은 하나 이상의 레이어를 구비하는 신경망일 수 있다.

[0044] 일 실시예에서, 제1 블록(Block-1)은 전처리 된 대상 이미지로부터 특징맵 A를 생성할 수 있다. 제2 블록(Block-2)은 특징맵 A를 입력 받아 특징맵 AB를 생성할 수 있다. 이러한 방식으로 제D 블록(Block-D)은 그 앞단의 블록으로부터 특징맵 AB를 입력 받아 특징맵 ABD를 생성할 수 잇다.

[0045] 출력부(102c)는 특징 추출부(102b)로부터 출력되는 특징에 기초하여 대상 이미지에 대한 예측 값을 출력할 수 있다. 여기서, 예측 값은 선생 모델(102)이 수행하고자 하는 태스크에 따라 달라질 수 있다. 일 실시예에서, 선생 모델(102)이 수행하고자 하는 태스크가 대상 이미지에 대한 객체 인식이라고 하면, 출력부(102c)는 대상 이미지 내 객체에 대한 분류 값을 출력할 수 있다. 이를 위해, 출력부(102c)는 글로벌 애버리지 풀링(global average pooling) 레이어 및 분류기(Ψ_T)를 포함할 수 있다.

[0046] 학생 모델(104)은 복수 개의 서브 학생 모델(104-1, 104-2, ~ 104-D)을 포함할 수 있다. 복수 개의 서브 학생 모델(104-1, 104-2, ~ 104-D)은 병렬로 배치될 수 있다. 복수 개의 서브 학생 모델은 선생 모델(102)에서 특징 추출부(102b)의 블록의 개수와 동일한 개수로 마련될 수 있다. 각 서브 학생 모델은 전처리부(111-1, 111-2, ~, 111-D), 특징 추출부(113-1, 113-2, ~, 113-D), 및 출력부(115-1, 115-2, ~, 115-D)를 포함할 수 있다. 복수 개의 서브 학생 모델(104-1, 104-2, ~ 104-D)은 각각 대상 이미지를 입력으로 받을 수 있다.

[0047] 복수 개의 서브 학생 모델(104-1, 104-2, ~ 104-D)의 특징 추출부(113-1, 113-2, ~, 113-D)는 선생 모델(102)의 특징 추출부(102b)의 복수 개의 블록 중 하나의 블록을 구비할 수 있다. 일 실시예서, 각 서브 학생 모델은 특징 추출부(102b)의 복수 개의 블록 중 그와 대응하는 순서의 블록을 구비할 수 있다. 이하에서는, 각 서브 학생 모델이 특징 추출부(102b)의 복수 개의 블록 중 그와 대응하는 순서의 블록을 구비하는 것을 일 실시예로 설명하지만, 이에 한정되는 것은 아니며 순서와 상관 없이 블록을 배치할 수도 있다.

[0048] 예를 들어, 제1 특징 추출부(113-1)는 특징 추출부(102b) 중 제1 블록(Block-1)을 구비할 수 있다. 제2 특징 추출부(113-2)는 특징 추출부(102b) 중 제2 블록(Block-2)을 구비할 수 있다. 제D 특징 추출부(113-D)는 특징 추

출부(102b) 중 제D 블록(Block-D)을 구비할 수 있다. 즉, 선생 모델(102)에서 복수 개의 블록(Block-1 ~ Block-D)은 순차적으로 연결(직렬로 연결)되었으나, 학생 모델(104)에서 복수 개의 블록(Block-1 ~ Block-D)은 병렬로 배치되게 된다.

[0049] 이때, 제1 특징 추출부(113-1)는 제1 전처리부(111-1)에 의해 전처리된 대상 이미지로부터 특징맵 A를 생성할 수 있다. 제2 특징 추출부(113-2)는 제2 전처리부(111-2)에 의해 전처리된 대상 이미지로부터 특징맵 B를 생성할 수 있다. 제D 특징 추출부(113-D)는 제D 전처리부(111-D)에 의해 전처리된 대상 이미지로부터 특징맵 D를 생성할 수 있다.

[0050] 여기서, 복수 개의 서브 학생 모델(104-1, 104-2, ~ 104-D)이 각각 대상 이미지를 입력 받기 때문에, 복수 개의 서브 학생 모델(104-1, 104-2, ~ 104-D)의 각 전처리부(111-1, 111-2, ~, 111-D)는 대상 이미지를 서로 다른 크기로 전처리 할 수 있다.

[0051] 제1 전처리부(111-1)는 선생 모델(102)의 전처리부(102a)와 동일한 크기로 대상 이미지를 전처리 할 수 있다. 제2 전처리부(111-2)는 제1 블록(Block-1)에서 출력되는 특징 맵의 크기와 동일한 크기가 되도록 대상 이미지를 전처리 할 수 있다.

[0052] 즉, 선생 모델(102)에서 제2 블록(Block-2)은 제1 블록(Block-1)에서 출력되는 특징 맵을 입력으로 하여 학습된 상태이므로, 제2 서브 학생 모델(104-2)에서 제2 전처리부(111-2)는 입력되는 대상 이미지를 제1 블록(Block-1)에서 출력되는 특징 맵의 크기와 동일한 크기가 되도록 전처리할 수 있다. 마찬가지로, 제D 전처리부(111-D)는 제2 블록(Block-2)에서 출력되는 특징 맵의 크기와 동일한 크기가 되도록 대상 이미지를 전처리 할 수 있다. 여기서, D는 3개인 것을 가정한 경우이다.

[0053] 예를 들어, 제1 전처리부(111-1)는 입력되는 대상 이미지를 32×32의 크기로 전처리 할 수 있다. 제2 전처리부(111-2)는 입력되는 대상 이미지를 16×16의 크기로 전처리 할 수 있다. 제D 전처리부(111-D)는 입력되는 대상 이미지를 8×8의 크기로 전처리 할 수 있다. 즉, 제1 전처리부(111-1)에서 제D 전처리부(111-D)로 갈수록 입력되는 대상 이미지를 좀더 작은 크기로 전처리 할 수 있다.

[0054] 복수 개의 서브 학생 모델(104-1, 104-2, ~ 104-D)의 출력부(115-1, 115-2, ~, 115-D)는 선생 모델(102)의 출력부(102c)와 대응하는 구조를 가질 수 있다. 각 출력부(115-1, 115-2, ~, 115-D)는 각 특징 추출부(113-1, 113-2, ~, 113-D)로부터 출력되는 특징 맵에 기초하여 대상 이미지에 대한 예측 값을 출력할 수 있다. 일 실시예에서, 각 출력부(115-1, 115-2, ~, 115-D)는 선생 모델(102)의 출력부(102c)와 마찬가지로 글로벌 애버리지 풀링(global average pooling) 레이어 및 분류기(Ψ_S^1, Ψ_S^2, ~, Ψ_S^D)를 각각 포함하여, 대상 이미지 내 객체에 대한 분류 값을 출력하도록 마련될 수 있다.

[0055] 한편, 학생 모델(104)은 선생 모델(102)로부터 지식 증류에 의한 학습을 받을 수 있다. 여기서, 복수 개의 서브 학생 모델(104-1, 104-2, ~ 104-D)의 특징 추출부(113-1, 113-2, ~, 113-D)는 선생 모델(102)의 특징 추출부(102b) 중 어느 하나의 블록만 포함하므로, 특징 추출 능력이 일부 손실되어 추상화된 특징을 추출하지 못하고 직접적인 특징만 생성하게 되며, 신경망의 깊이가 감소함에 따라 유도 편향이 약해지는 문제가 발생할 수 있게 된다.

[0056] 이에 개시되는 실시예에서는, 선생 모델(102)의 출력부(102c)의 예측 값과 학생 모델(104)의 출력부의 예측 값 간에 손실 함수 및 학생 모델(104)의 출력부의 예측 값과 정답 값 간의 손실 함수를 산출할 뿐 아니라, 선생 모델(102)의 각 블록에서 생성한 특징맵과 학생 모델(104)의 각 블록에서 생성한 특징맵 간에 손실 함수를 산출하여 이러한 문제를 해결할 수 있다.

[0057] 구체적으로, 학생 모델(104)의 학습은 각 서브 학생 모델(104-1, 104-2, ~ 104-D)과 선생 모델(102) 간에 이루어질 수 있다. 각 서브 학생 모델(104-1, 104-2, ~ 104-D)은 3개의 손실 함수를 목적 함수로 할 수 있다.

[0058] 각 서브 학생 모델은 해당 서브 학생 모델의 출력부에서 출력하는 예측 값과 정답 값(ground truth) 간의 손실 함수인 제1 손실 함수를 가질 수 있다. 제1 손실 함수는 서브 학생 모델의 출력부에서 출력하는 예측 값과 정답 값(ground truth) 간의 크로스 엔트로피 손실(cross-entropy loss)일 수 있다.

[0059] 각 서브 학생 모델은 해당 출력부에서 출력하는 예측 값과 선생 모델(102)의 출력부(102c)에서 출력하는 예측 값 간의 손실 함수인 제2 손실 함수를 가질 수 있다. 또한, 각 서브 학생 모델은 해당 특징 추출부의 블록에서 생성한 특징맵과 선생 모델(102)의 특징 추출부(102b)의 그에 대응하는 블록에서 생성한 특징맵 간의 손실 함수

인 제3 손실 함수를 가질 수 있다.

[0060] 도 2는 본 발명의 일 실시예에서 제1 서브 학생 모델(104-1)과 선생 모델(102) 간의 제2 손실 함수 및 제3 손실 함수를 나타낸 도면이다. 도 2를 참조하면, 제1 서브 학생 모델(104-1)은 제1 출력부(115-1)에서 출력하는 예측 값과 선생 모델(102)의 출력부(102c)에서 출력하는 예측 값 간의 제2 손실 함수를 가질 수 있다.

[0061] 일 실시예에서, 제2 손실 함수는 로짓 증류(logit distillation)에 의한 손실일 수 있다. 로짓 증류(logit distillation)는 선생 모델(102)이 예측한 예측 값의 소프트라벨(각 클래스가 정답일 확률)과 제1 서브 학생 모델(104-1)이 예측한 예측 값의 소프트라벨 간의 차이가 최소화 되도록 학습하는 방식일 수 있다. 여기서, 소프트라벨에 의한 확률 분포는 아래 수학식 1과 같이 나타낼 수 있다.

[0062] (수학식 1)

[0063] $$p^i(x;\tau) = softmax(\phi(x);\tau) = \frac{e^{\phi_i(x)/\tau}}{\sum_k e^{\phi_k(x)/\tau}}$$

[0064] x : 입력

[0065] Ψ_i : i번째 로짓(logit)

[0066] k : 로짓의 전체 개수

[0067] τ : 온도(temperature)

[0068] 수학식 1에서, 분자는 전체 로짓 행렬 중 i번째 로짓의 값이 활성화된 값을 의미하고, 분모는 로짓 행렬의 원소들이 활성화된 값의 합을 의미할 수 있다.

[0069] 그리고, 로짓 증류(logit distillation)에 의한 제2 손실 함수($\mathcal{L}_{kd}$)는 하기 수학식 2로 나타낼 수 있다.

[0070] (수학식 2)

[0071] $$\mathcal{L}_{kd} = -\tau^2 \sum_{x\sim\mathcal{D}_x} \sum_{i=1}^{C} p^i_{\phi_T}(x;\tau) log(p^i_{\phi_L}(x;\tau))$$

[0072] Ψ_T : 선생 모델

[0073] Ψ_L : 서브 학생 모델

[0074] C : 클래스의 개수

[0075] D_x : 입력 데이터 셋의 개수

[0076] 수학식 2에서 $p^i_{\phi_T}(x;\tau)$는 선생 모델의 소프트라벨 값을 의미하고, $p^i_{\phi_L}(x;\tau)$은 서브 학생 모델의 소프트라벨 값을 의미한다.

[0077] 또한, 제1 서브 학생 모델(104-1)은 제1 특징 추출부(113-1)의 블록에서 생성한 특징맵(F_i^S)과 선생 모델(102)의 특징 추출부(102b)의 제1 블록에서 생성한 특징맵(F_i^T) 간의 제3 손실 함수를 가질 수 있다. 일 실시예에서, 제3 손실 함수는 특징맵(F_i^S)과 특징맵(F_i^T) 간의 거리가 최소화되도록 나타낼 수 있다. 일 실시예에서, 제3 손실 함수($\mathcal{L}_{feature}$)는 하기 수학식 3으로 나타낼 수 있다.

[0078] (수학식 3)

[0079] $$\mathcal{L}_{feature}(\mathcal{F}_i^T, \mathcal{F}_i^S) = \|\mathcal{F}_i^T - \mathcal{F}_i^S\|_2^2$$

[0080] F_i^T : 선생 모델의 특징 추출부의 i번째 블록에서 생성한 특징맵

[0081] F_i^S : i번째 서브 학생 모델의 특징 추출부의 블록에서 생성한 특징맵

[0082] 그러면, 각 서브 학생 모델을 학습하기 위한 전체 손실 함수는 하기 수학식 4로 나타낼 수 있다.

[0083] (수학식 4)

[0084] $$\mathcal{L} = \lambda_1 \mathcal{L}_{ce} + \lambda_2 \mathcal{L}_{kd} + \mathcal{L}_{feature}$$

[0085] $\mathcal{L}_{ce}$: 제1 손실 함수

[0086] $\mathcal{L}_{kd}$: 제2 손실 함수

[0087] $\mathcal{L}_{feature}$: 제3 손실 함수

[0088] λ_1, λ_2 : 기 설정된 가중치

[0089] λ_1, λ_2는 밸런싱 하이퍼파라미터(balancing hyperparameter)이다. 일 실시예에서, λ_1, λ_2는 각각 1, 0.8일 수 있으나, 이에 한정되는 것은 아니다.

[0090] 한편, 각 서브 학생 모델(104-1, 104-2, ~ 104-D)의 학습이 완료된 경우, 추론 단계에서 대상 이미지가 각 서브 학생 모델(104-1, 104-2, ~ 104-D)되면 각 서브 학생 모델(104-1, 104-2, ~ 104-D)에서 출력하는 예측 값에 기초하여 최종 예측 값을 산출할 수 있다.

[0091] 즉, 각 서브 학생 모델(104-1, 104-2, ~ 104-D)에서 출력하는 예측 값을 앙상블 하여 최종 예측 값을 산출할 수 있다. 일 실시예에서, 각 서브 학생 모델(104-1, 104-2, ~ 104-D)에서 출력하는 예측 값을 평균하거나 대표 값을 선출하거나 각 가중치를 부여하는 등의 방식을 통해 최종 예측 값을 산출할 수 있다.

[0093] 도 3은 본 발명의 일 실시예에서 선생 모델(102)의 전체 실행 시간과 학생 모델(104)의 전체 실행 시간을 비교한 도면이다. 도 3의 (a)를 참조하면, 선생 모델(102)의 전체 실행 시간($t_{\Psi T}$)은 하기 수학식 5로 나타낼 수 있다.

[0094] (수학식 5)

[0095] $$t_{\phi_T} = t_p + \sum_{i=1}^{D} t_i$$

[0096] t_p : 대상 이미지를 전처리 하는데 소요되는 시간

[0097] t_i : 특징 추출부의 i번째 블록의 실행 시간

[0098] D : 특징 추출부의 블록의 전체 개수

[0099] 즉, 선생 모델(102)의 경우, 전체 실행 시간($t_{\Psi T}$)은 대상 이미지를 전처리 하는 데 소요되는 시간에 각 블록(Block-1 ~ Block-D)의 실행시간의 총합을 더한 값이 된다.

[0100] 반면, 도 3의 (b)를 참조하면, 학생 모델(104)의 전체 실행 시간($t_{\Psi T}$)은 하기 수학식 6으로 나타낼 수 있다.

[0101] (수학식 6)

[0102] $$t_{\phi_T} = t_p + max(t_i)$$

[0103] t_p : 대상 이미지를 전처리 하는데 소요되는 시간

[0104] $max(t_i)$: i번째 블록의 실행 시간 중 최대값

[0105] 즉, 학생 모델(104)의 경우, 전체 실행 시간($t_{\Psi T}$)은 대상 이미지를 전처리 하는 데 소요되는 시간에 각 특징 추출부의 블록의 실행 시간 중 최대값을 더한 값이 된다.

[0106] 여기서, 대상 이미지를 전처리 하는 데 소요되는 시간이 선생 모델(102) 및 학생 모델(104)에서 동일하다고 가정하면, 선생 모델(102)은 특징 추출부(102b)에서 복수 개의 블록이 순차적으로 연결된 관계로 각 블록의 실행 시간의 총합이 전체 실행 시간이 되는 반면, 학생 모델(104)은 각 특징 추출부(113-1, 113-2, ~, 113-D)가 병렬로 배치되는 바 각 특징 추출부(113-1, 113-2, ~, 113-D)의 블록 중 최대 실행 시간이 전체 실행 시간이 되게 된다. 따라서, 학생 모델(104)은 선생 모델(102)에 비하여 인공 신경망의 크기가 작을 뿐만 아니라, 태스크의 실행 시간도 효과적으로 단출할 수 있게 된다.

[0108] 도 4는 예시적인 실시예들에서 사용되기에 적합한 컴퓨팅 장치를 포함하는 컴퓨팅 환경(10)을 예시하여 설명하기 위한 블록도이다. 도시된 실시예에서, 각 컴포넌트들은 이하에 기술된 것 이외에 상이한 기능 및 능력을 가질 수 있고, 이하에 기술된 것 이외에도 추가적인 컴포넌트를 포함할 수 있다.

[0109] 도시된 컴퓨팅 환경(10)은 컴퓨팅 장치(12)를 포함한다. 일 실시예에서, 컴퓨팅 장치(12)는 인공 신경망 아키텍처(100)를 구비하는 장치일 수 있다. 또한, 컴퓨팅 장치(12)는 인공 신경망 아키텍처(100)를 학습하기 위한 장치일 수 있다.

[0110] 컴퓨팅 장치(12)는 적어도 하나의 프로세서(14), 컴퓨터 판독 가능 저장 매체(16) 및 통신 버스(18)를 포함한다. 프로세서(14)는 컴퓨팅 장치(12)로 하여금 앞서 언급된 예시적인 실시예에 따라 동작하도록 할 수 있다. 예컨대, 프로세서(14)는 컴퓨터 판독 가능 저장 매체(16)에 저장된 하나 이상의 프로그램들을 실행할 수 있다. 상기 하나 이상의 프로그램들은 하나 이상의 컴퓨터 실행 가능 명령어를 포함할 수 있으며, 상기 컴퓨터 실행 가능 명령어는 프로세서(14)에 의해 실행되는 경우 컴퓨팅 장치(12)로 하여금 예시적인 실시예에 따른 동작들을 수행하도록 구성될 수 있다.

[0111] 컴퓨터 판독 가능 저장 매체(16)는 컴퓨터 실행 가능 명령어 내지 프로그램 코드, 프로그램 데이터 및/또는 다른 적합한 형태의 정보를 저장하도록 구성된다. 컴퓨터 판독 가능 저장 매체(16)에 저장된 프로그램(20)은 프로세서(14)에 의해 실행 가능한 명령어의 집합을 포함한다. 일 실시예에서, 컴퓨터 판독 가능 저장 매체(16)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 컴퓨팅 장치(12)에 의해 액세스되고 원하는 정보를 저장할 수 있는 다른 형태의 저장 매체, 또는 이들의 적합한 조합일 수 있다.

[0112] 통신 버스(18)는 프로세서(14), 컴퓨터 판독 가능 저장 매체(16)를 포함하여 컴퓨팅 장치(12)의 다른 다양한 컴포넌트들을 상호 연결한다.

[0113] 컴퓨팅 장치(12)는 또한 하나 이상의 입출력 장치(24)를 위한 인터페이스를 제공하는 하나 이상의 입출력 인터페이스(22) 및 하나 이상의 네트워크 통신 인터페이스(26)를 포함할 수 있다. 입출력 인터페이스(22) 및 네트워크 통신 인터페이스(26)는 통신 버스(18)에 연결된다. 입출력 장치(24)는 입출력 인터페이스(22)를 통해 컴퓨팅 장치(12)의 다른 컴포넌트들에 연결될 수 있다. 예시적인 입출력 장치(24)는 포인팅 장치(마우스 또는 트랙패드 등), 키보드, 터치 입력 장치(터치패드 또는 터치스크린 등), 음성 또는 소리 입력 장치, 다양한 종류의 센서 장치 및/또는 촬영 장치와 같은 입력 장치, 및/또는 디스플레이 장치, 프린터, 스피커 및/또는 네트워크 카드와 같은 출력 장치를 포함할 수 있다. 예시적인 입출력 장치(24)는 컴퓨팅 장치(12)를 구성하는 일 컴포넌트로서 컴퓨팅 장치(12)의 내부에 포함될 수도 있고, 컴퓨팅 장치(12)와는 구별되는 별개의 장치로 컴퓨팅 장치(12)와 연결될 수도 있다.

[0115] 도 5는 본 발명의 일 실시예에 따른 인공 신경망의 학습 방법을 나타낸 흐름도이다. 도시된 흐름도에서는 상기 방법을 복수 개의 단계로 나누어 기재하였으나, 적어도 일부의 단계들은 순서를 바꾸어 수행되거나, 다른 단계와 결합되어 함께 수행되거나, 생략되거나, 세부 단계들로 나뉘어 수행되거나, 또는 도시되지 않은 하나 이상의 단계가 부가되어 수행될 수 있다.

[0116] 도 5를 참조하면, 컴퓨팅 장치(12)는 선생 모델(102)을 학습할 수 있다(S 101). 일 실시예에서, 선생 모델(102)은 학생 모델(104)이 수행할 태스크에 대해 기 설정된 학습 데이터 셋을 이용하여 학습될 수 있다.

[0117] 다음으로, 컴퓨팅 장치(12)는 선생 모델(102)의 특징 추출부(102b)를 복수 개의 블록으로 분할 할 수 있다(S 103). 즉, 컴퓨팅 장치(12)는 특징 추출부(102b)의 순차적으로 연결된 복수 개의 블록을 각각 별개의 블록으로 분할 할 수 있다.

[0118] 다음으로, 컴퓨팅 장치(12)는 분할된 복수 개의 블록에 기초하여 복수 개의 서브 학생 모델들을 병렬로 배치할 수 있다(S 105). 컴퓨팅 장치(12)는 분할된 복수 개의 블록을 순서에 따라 각 서브 학생 모델의 특징 추출부에 배치할 수 있다.

[0119] 다음으로, 컴퓨팅 장치(12)는 각 서브 학생 모델의 특징 추출부의 전단에 전처리부를 배치할 수 있다(S 107). 이때, 컴퓨팅 장치(12)는 제1 전처리부(111-1)는 선생 모델(102)의 전처리부(102a)와 동일한 크기로 대상 이미지를 전처리하도록 할 수 있다. 또한, 컴퓨팅 장치(12)는 제n 전처리부(111-n)(학생 모델에서 n번째 전처리부)는 선생 모델(102)에서 n-1번째 블록에서 출력되는 특징 맵의 크기와 동일한 크기로 대상 이미지를 전처리하도록 할 수 있다.

[0120] 다음으로, 컴퓨팅 장치(12)는 각 서브 학생 모델의 특징 추출부의 후단에 출력부를 배치할 수 있다(S 109). 서브 학생 모델의 각 출력부는 해당 특징 추출부로부터 출력되는 특징 맵에 기초하여 대상 이미지에 대한 예측 값을 출력하도록 마련될 수 있다.

[0121] 다음으로, 컴퓨팅 장치(12)는 각 서브 학생 모델을 학습할 수 있다(S 111). 컴퓨팅 장치(12)는 각 서브 학생 모델이 제1 손실 함수, 제2 손실 함수, 및 제3 손실 함수를 최소화하도록 학습할 수 있다.

[0123] 이상에서 본 발명의 대표적인 실시예들을 상세하게 설명하였으나, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자는 상술한 실시예에 대하여 본 발명의 범주에서 벗어나지 않는 한도 내에서 다양한 변형이 가능함을 이해할 것이다. 그러므로 본 발명의 권리범위는 설명된 실시예에 국한되어 정해져서는 안 되며, 후술하는 특허청구범위뿐만 아니라 이 특허청구범위와 균등한 것들에 의해 정해져야 한다.

부호의 설명

[0125] 10 : 컴퓨팅 환경

12 : 컴퓨팅 장치

14 : 프로세서

16 : 컴퓨터 판독 가능 저장 매체

18 : 통신 버스

20 : 프로그램

22 : 입출력 인터페이스

24 : 입출력 장치

26 : 네트워크 통신 인터페이스

100 : 인공 신경망 아키텍처

102 : 선생 모델

102a : 전처리부

102b : 특징 추출부

102c : 출력부

104 : 학생 모델

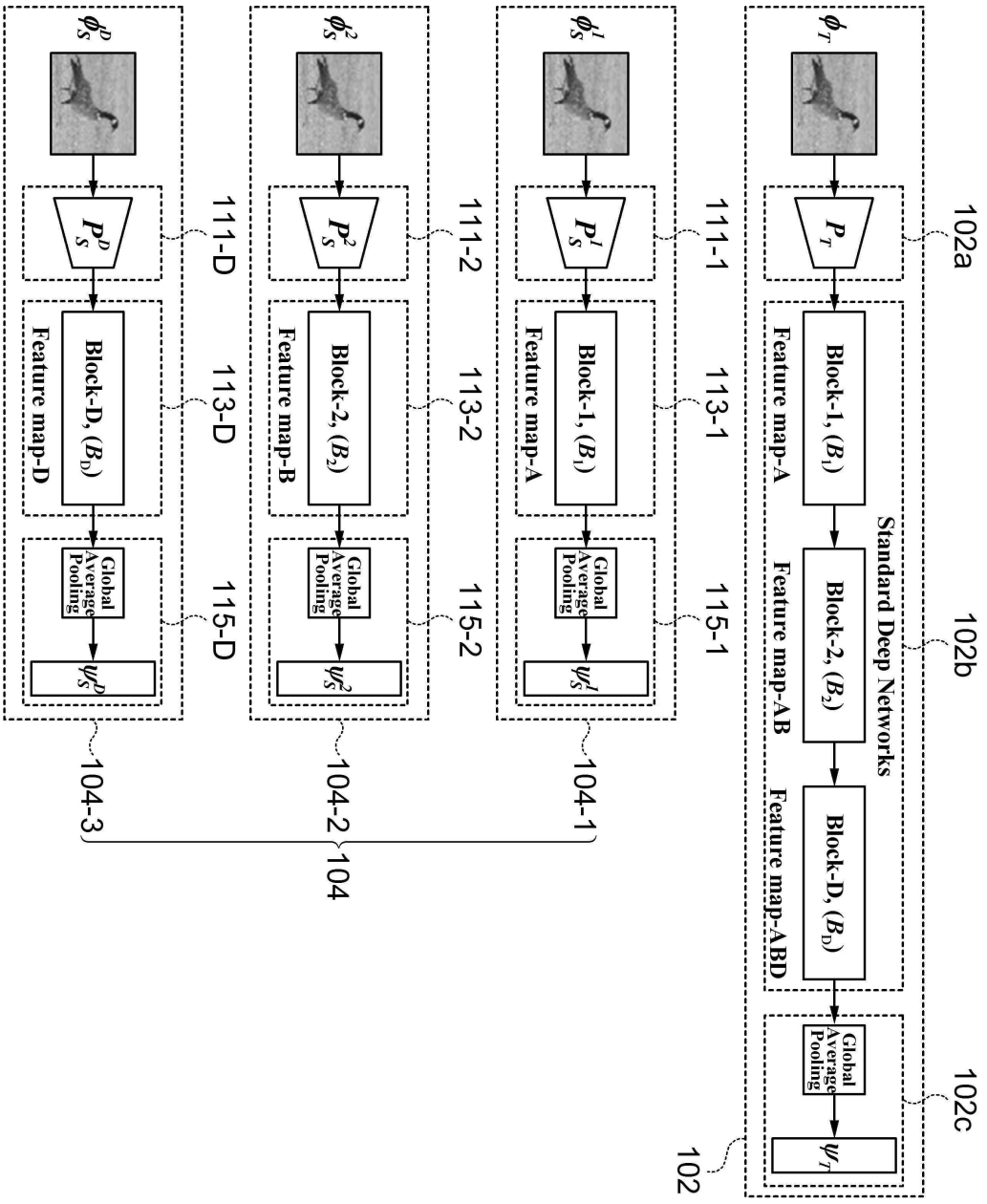
111 : 전처리부

113 : 특징 추출부

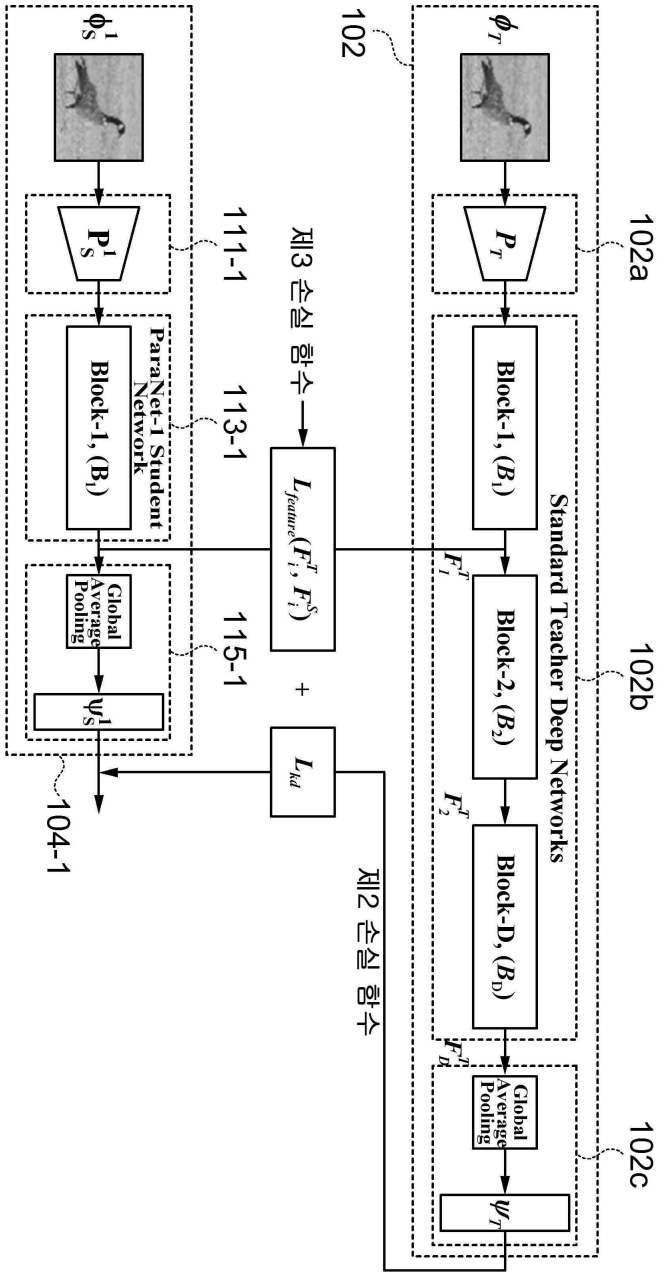
115 : 출력부

도면

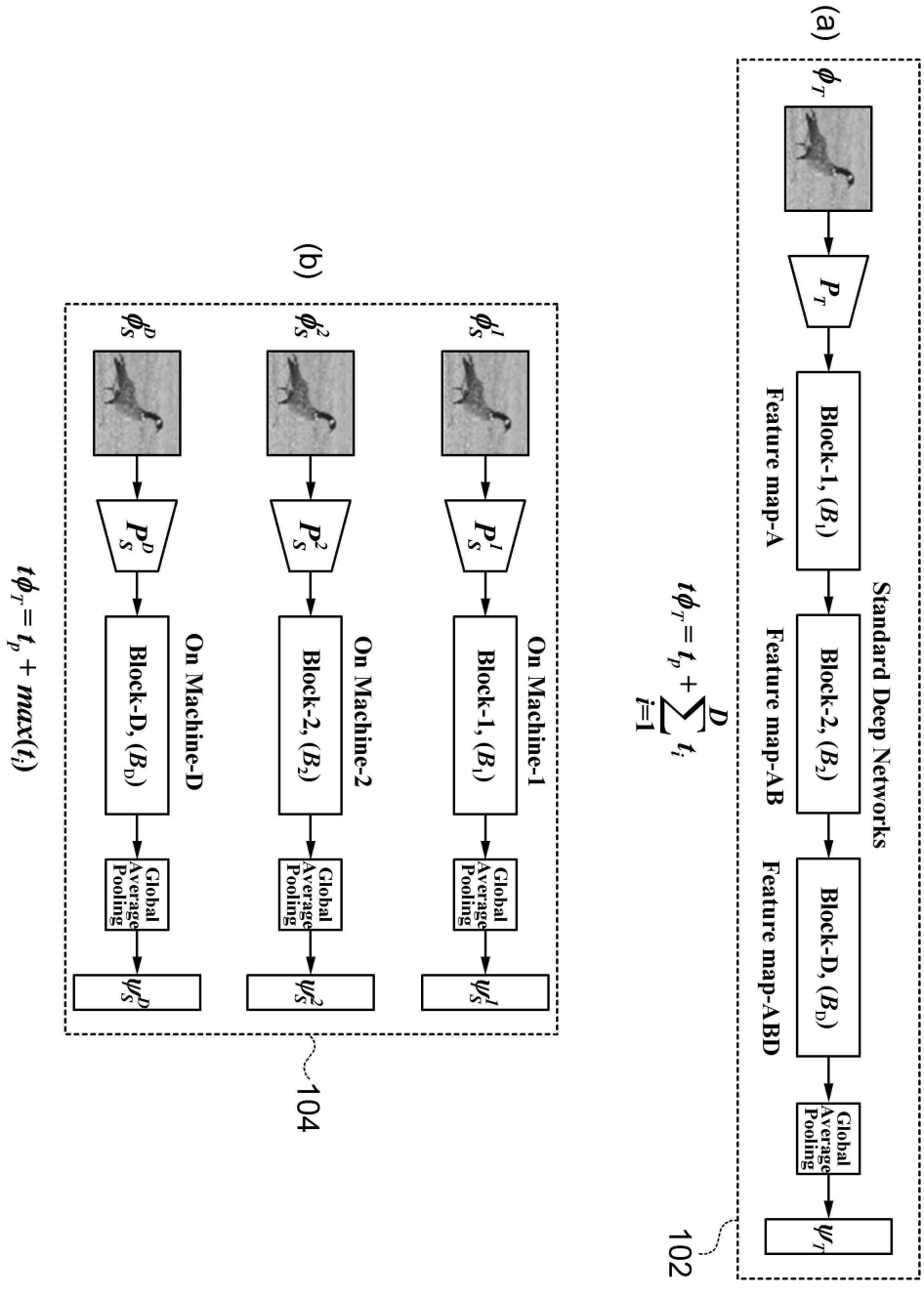
도면1



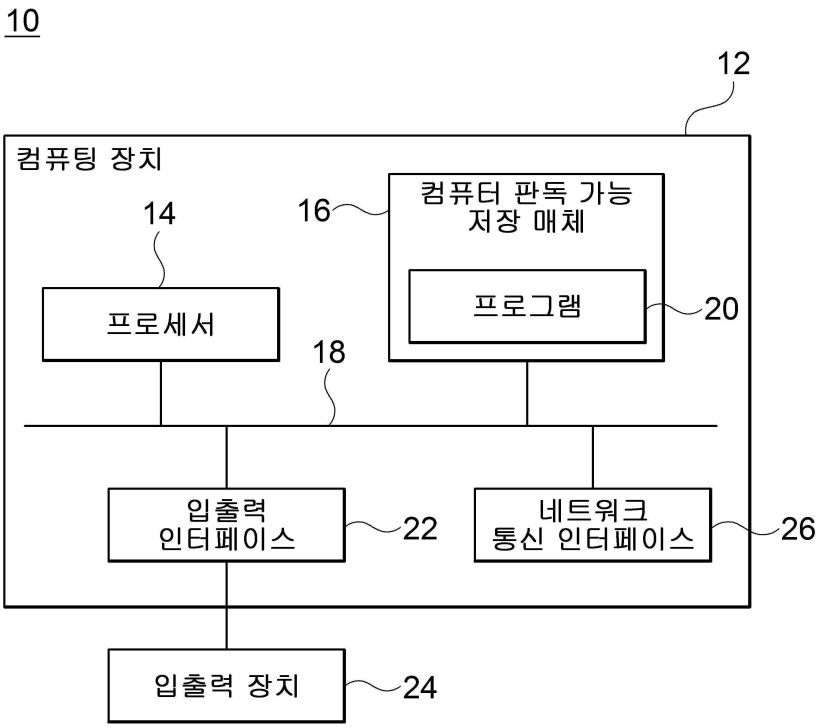
도면2


102
102a
102b
102c
Standard Teacher Deep Networks
Block-1, (B1)
Block-2, (B2)
Block-D, (BD)
Global Average Pooling
111-1
113-1
115-1
104-1
ParaNet-1 Student Network
제3 손실 함수
제2 손실 함수
+

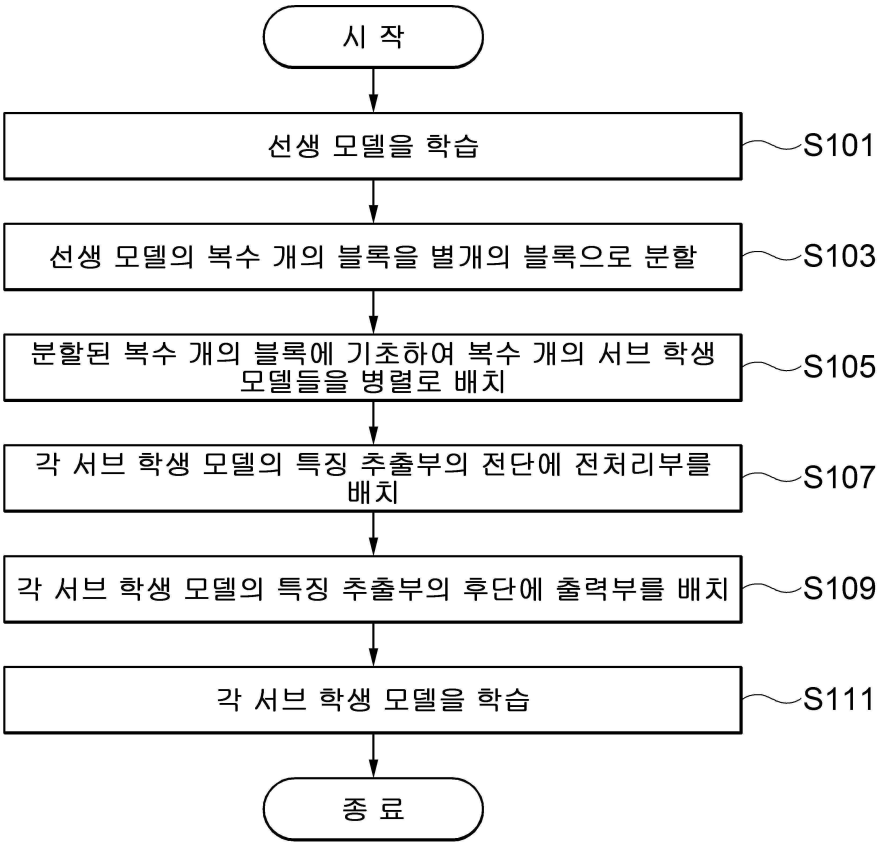
도면3


(a)
Standard Deep Networks
Feature map-A
Feature map-AB
Feature map-ABD
Block-1, (B1)
Block-2, (B2)
Block-D, (BD)
Global Average Pooling
102
(b)
On Machine-1
On Machine-2
On Machine-D
104

도면4


10
12
컴퓨팅 장치
14
프로세서
16
컴퓨터 판독 가능
저장 매체
프로그램
20
18
입출력
인터페이스
22
네트워크
통신 인터페이스
26
입출력 장치
24

도면5


시 작
선생 모델을 학습
S101
선생 모델의 복수 개의 블록을 별개의 블록으로 분할
S103
분할된 복수 개의 블록에 기초하여 복수 개의 서브 학생
모델들을 병렬로 배치
S105
각 서브 학생 모델의 특징 추출부의 전단에 전처리부를
배치
S107
각 서브 학생 모델의 특징 추출부의 후단에 출력부를 배치
S109
각 서브 학생 모델을 학습
S111
종 료