

(12) 특허협력조약에 의하여 공개된 국제출원

(19) 세계지식재산권기구
국제사무국

(43) 국제공개일
2021년 9월 16일 (16.09.2021) WIPO | PCT



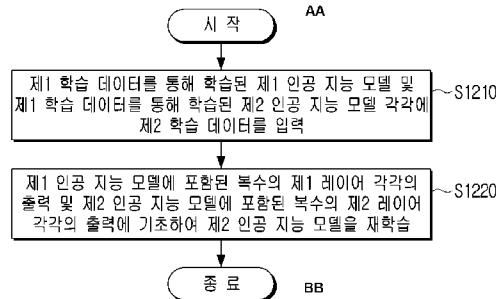
(10) 국제공개번호
WO 2021/182748 A1

- (51) 국제특허분류: G06N 3/04 (2006.01) G06N 20/00 (2019.01) EMIC COOPERATION FOUNDATION) [KR/KR]; 16499 경기도 수원시 영통구 월드컵로 206, Gyeonggi-do (KR).
G06N 3/08 (2006.01) G06N 5/02 (2006.01)
- (21) 국제출원번호: PCT/KR2021/000891 (72) 발명자: 송원석 (SONG, Wonseok); 16677 경기도 수원시 영통구 삼성로 129, Gyeonggi-do (KR). 손원철 (SON, Wonchul); 16499 경기도 수원시 영통구 월드컵로 206, Gyeonggi-do (KR). 황원준 (HWANG, Wonjun); 16499 경기도 수원시 영통구 월드컵로 206, Gyeonggi-do (KR). 구강모 (KOO, Gangmo); 16677 경기도 수원시 영통구 삼성로 129, Gyeonggi-do (KR). 김영빈 (KIM, Youngbin); 16499 경기도 수원시 영통구 월드컵로 206, Gyeonggi-do (KR). 문영수 (MOON, Youngsu); 16677 경기도 수원시 영통구 삼성로 129, Gyeonggi-do (KR). 이희중 (LEE, Huijung); 16677 경기도 수원시 영통구 삼성로 129, Gyeonggi-do (KR). 조영훈 (JO, Younghun); 16677 경기도 수원시 영통구 삼성로 129, Gyeonggi-do (KR).
- (22) 국제출원일: 2021년 1월 22일 (22.01.2021)
- (25) 출원언어: 한국어
- (26) 공개언어: 한국어
- (30) 우선권정보: 10-2020-0029849 2020년 3월 10일 (10.03.2020) KR
10-2020-0114122 2020년 9월 7일 (07.09.2020) KR
- (71) 출원인: 삼성전자주식회사 (SAMSUNG ELECTRONICS CO., LTD.) [KR/KR]; 16677 경기도 수원시 영통구 삼성로 129, Gyeonggi-do (KR). 아주대학교 산학협력단 (AJOU UNIVERSITY INDUSTRY-ACAD-

(54) Title: ELECTRONIC DEVICE AND CONTROL METHOD THEREOF

(54) 발명의 명칭: 전자 장치 및 그 제어 방법

[도.12]



S1210 ... Input second learning data to each of first artificial intelligence model learned through first learning data and second artificial intelligence model learned through first learning data

S1220 ... Relearn second artificial intelligence model on basis of output of each of plurality of first layers included in first artificial intelligence model and output of each of plurality of second layers included in second artificial intelligence model

AA ... Start

BB ... End

(57) Abstract: An electronic device is disclosed. The present electronic device may comprise: a memory in which information on a first artificial intelligence model learned through first learning data and information on a second artificial intelligence model learned through the first learning data are stored; and a processor connected to the memory to control the electronic device, wherein the processor inputs second learning data to each of the first artificial intelligence model and the second artificial intelligence model and relearns the second artificial intelligence model on the basis of an output of each of a plurality of first layers included in the first artificial intelligence model and an output of each of a plurality of second layers included in the second artificial intelligence model, each of the plurality of first layers includes a plurality of two-dimensional filters, and each of the plurality of second layers includes a plurality of filters obtained

[다음 쪽 계속]

(74) 대리인: 김태헌 등 (KIM, Tae-hun et al.); 06626 서울시
서초구 강남대로343 신덕빌딩 9층, Seoul (KR).

(81) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의 국내 권리의 보호를 위하여): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의 국내 권리의 보호를 위하여): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 유라시아 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 유럽 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

공개:

— 국제조사보고서와 함께 (조약 제21조(3))

by reducing the size of each of the plurality of two-dimensional filters of a corresponding first layer.

(57) 요약서: 전자 장치가 개시된다. 본 전자 장치는 제1 학습 데이터를 통해 학습된 제1 인공 지능 모델에 대한 정보 및 제1 학습 데이터를 통해 학습된 제2 인공 지능 모델에 대한 정보가 저장된 메모리 및 메모리와 연결되어 전자 장치를 제어하는 프로세서를 포함하며, 프로세서는 제2 학습 데이터를 제1 인공 지능 모델 및 제2 인공 지능 모델 각각에 입력하고, 제1 인공 지능 모델에 포함된 복수의 제1 레이어 각각의 출력 및 제2 인공 지능 모델에 포함된 복수의 제2 레이어 각각의 출력에 기초하여 제2 인공 지능 모델을 재학습하며, 복수의 제1 레이어 각각은 복수의 2차원 필터를 포함하고, 복수의 제2 레이어 각각은 대응되는 제1 레이어의 복수의 2차원 필터 각각의 크기가 축소된 복수의 필터를 포함할 수 있다.

명세서

발명의 명칭: 전자 장치 및 그 제어 방법

기술분야

- [1] 본 개시는 전자 장치 및 그 제어 방법에 대한 것으로, 더욱 상세하게는 인공지능 모델(Artificial Intelligence Model)을 학습하는 전자 장치 및 그 제어 방법에 대한 것이다.

배경기술

- [2] 인간 수준의 지능이 구현된 인공지능 모델이 다양한 분야에서 이용되고 있다. 인공지능 모델은 기존의 룰(rule) 기반 모델과는 달리 기계가 스스로 학습하고 판단하며 똑똑해지는 모델이며, 기존 룰 기반 모델은 점차 딥러닝 기반 인공지능 모델로 대체되고 있다.
- [3] 인공지능 모델은 일반적으로 용량이 클수록 성능이 향상된다. 다만, 전자 장치의 소형화에 따라 인공지능 모델의 성능을 유지하면서도 용량을 축소시킬 필요가 있다. 이를 위한 다양한 방법이 개발되고 있으며, 그 중 Knowledge distillation은 용량이 큰 네트워크(Teacher network)의 지식을 실제로 사용하고자 하는 용량이 작은 네트워크(Student network)에게 전달하는 방법을 이용하고 있다.
- [4] 도 1a에 도시된 바와 같이, Knowledge distillation은 용량이 작은 네트워크(Student network)의 출력이 용량이 큰 네트워크(Teacher network)의 출력과 유사해지도록 용량이 작은 네트워크(Student network)를 재학습하는 방법이다.
- [5] 다만, 이 방법은 각 네트워크의 마지막 softmax 정보만을 이용하기 때문에 하이-레벨(high-level)의 특징만이 부각되고 로우-레벨(low-level)의 특징은 부각되지 않는 문제가 있다. 즉, 객체의 특징 추출이 중요한 응용에서는 해당 방법의 적용이 어려운 문제가 있다.
- [6] 또는, 도 1b에 도시된 바와 같이, 각 레이어의 출력을 비교하는 방법도 있을 수 있으나, 각 레이어의 출력의 크기가 달라 어느 하나의 크기를 변경해줄 필요가 있다. 이 경우, 학습 과정에서 추가 연산이 필요하며, 레이어의 개수가 다르거나 레이어에 포함된 채널의 개수가 다른 경우 로우-레벨의 특징이 정확히 반영되지 않는 문제도 있다.
- [7] 그에 따라, 로우-레벨의 특징을 좀더 정확히 반영하면서도 성능이 개선된 학습 방법이 개발될 필요가 있다.

발명의 상세한 설명

기술적 과제

- [8] 본 개시는 상술한 필요성에 따른 것으로, 본 개시의 목적은 Knowledge Distillation을 통해 인공지능 모델을 압축하면서도 성능을 향상시키기 위한 전자

장치 및 그 제어 방법을 제공함에 있다.

기술적 해결방법

- [9] 이상과 같은 목적을 달성하기 위한 본 개시의 일 실시 예에 따르면, 전자 장치는 제1 학습 데이터를 통해 학습된 제1 인공 지능 모델에 대한 정보 및 상기 제1 학습 데이터를 통해 학습된 제2 인공 지능 모델에 대한 정보가 저장된 메모리 및 상기 메모리와 연결되어 상기 전자 장치를 제어하는 프로세서를 포함하며, 상기 프로세서는 제2 학습 데이터를 상기 제1 인공 지능 모델 및 상기 제2 인공 지능 모델 각각에 입력하고, 상기 제1 인공 지능 모델에 포함된 복수의 제1 레이어 각각의 출력 및 상기 제2 인공 지능 모델에 포함된 복수의 제2 레이어 각각의 출력에 기초하여 상기 제2 인공 지능 모델을 재학습하며, 상기 복수의 제1 레이어 각각은 복수의 2차원 필터를 포함하고, 상기 복수의 제2 레이어 각각은 대응되는 제1 레이어의 복수의 2차원 필터 각각의 크기가 축소된 복수의 필터를 포함할 수 있다.
- [10] 또한, 상기 복수의 제1 레이어의 개수는 상기 복수의 제2 레이어의 개수와 동일하고, 상기 프로세서는 상기 복수의 제1 레이어 각각의 출력을 대응되는 제2 레이어의 출력과 비교하여 복수의 비교 결과를 획득하고, 상기 복수의 비교 결과에 기초하여 상기 제2 인공 지능 모델을 재학습할 수 있다.
- [11] 그리고, 상기 복수의 제1 레이어 각각에 포함된 상기 복수의 2차원 필터의 개수는 대응되는 제2 레이어에 포함된 상기 복수의 필터의 개수와 동일하며, 상기 복수의 제1 레이어 각각의 출력의 크기는 대응되는 제2 레이어의 출력의 크기와 동일할 수 있다.
- [12] 또한, 상기 메모리는 상기 복수의 제1 레이어 각각의 출력 및 상기 복수의 제2 레이어 각각의 출력을 구별하도록 학습된 제3 인공 지능 모델에 대한 정보를 저장하며, 상기 프로세서는 상기 복수의 제2 레이어 각각의 출력을 상기 제3 인공 지능 모델 각각에 입력하고, 상기 제3 인공 지능 모델로부터 출력되는 상기 복수의 제2 레이어의 출력에 대한 복수의 구별 결과를 획득하며, 상기 복수의 비교 결과 및 상기 복수의 구별 결과에 기초하여 상기 제2 인공 지능 모델을 재학습할 수 있다.
- [13] 그리고, 상기 프로세서는 상기 제2 학습 데이터를 상기 재학습된 제2 인공 지능 모델 각각에 입력하고, 상기 복수의 제1 레이어 각각의 출력 및 상기 재학습된 제2 인공 지능 모델에 포함된 복수의 제3 레이어 각각의 출력을 구별하도록 상기 제3 인공 지능 모델을 재학습할 수 있다.
- [14] 또한, 상기 프로세서는 상기 복수의 비교 결과 및 상기 복수의 구별 결과를 가중합하여 상기 제2 인공 지능 모델을 재학습할 수 있다.
- [15] 그리고, 상기 프로세서는 상기 복수의 제1 레이어 중 마지막 레이어에 대응되는 비교 결과의 가중치를 나머지 레이어에 대응되는 복수의 비교 결과의 가중치 이상으로 부여하여 상기 제2 인공 지능 모델을 재학습할 수 있다.

- [16] 또한, 상기 복수의 필터 각각은 $1 \times N$ 형태 또는 $N \times 1$ 형태이고, 상기 프로세서는 상기 메모리로부터 입력 데이터를 행(row) 단위 또는 컬럼(column) 단위로 독출하고, 상기 독출된 입력 데이터를 상기 재학습된 제2 인공 지능 모델에 입력하여 상기 입력 데이터를 처리할 수 있다.
- [17] 그리고, 상기 제1 인공 지능 모델 및 상기 제2 인공 지능 모델은 컨볼루션 신경망(Convolutional Neural Network, CNN) 형태로 구현될 수 있다.
- [18] 또한, 상기 메모리는 상기 제1 학습 데이터를 통해 학습된 보조 인공 지능 모델에 대한 정보를 더 저장하고, 상기 프로세서는 상기 제2 학습 데이터를 상기 보조 인공 지능 모델에 입력하고, 상기 복수의 제1 레이어 각각의 출력 및 상기 보조 인공 지능 모델에 포함된 복수의 보조 레이어 각각의 출력에 기초하여 상기 보조 인공 지능 모델을 재학습하며, 상기 제2 학습 데이터를 상기 재학습된 보조 인공 지능 모델에 입력하고, 상기 재학습된 보조 인공 지능 모델에 포함된 복수의 보조 레이어 각각의 출력 및 상기 복수의 제2 레이어 각각의 출력에 기초하여 상기 제2 인공 지능 모델을 재학습할 수 있다. 여기서, 상기 복수의 보조 레이어 각각은 대응되는 제1 레이어의 복수의 2차원 필터 각각의 크기가 축소된 복수의 필터를 포함할 수 있다.
- [19] 그리고, 상기 메모리는 상기 제1 학습 데이터를 통해 학습된 제1 보조 인공 지능 모델에 대한 정보 및 상기 제1 학습 데이터를 통해 학습된 제2 보조 인공 지능 모델에 대한 정보를 더 저장하고, 상기 프로세서는 상기 제2 학습 데이터를 상기 제1 보조 인공 지능 모델 및 상기 제2 보조 인공 지능 모델에 입력하고, 상기 복수의 제1 레이어 각각의 출력 및 상기 제1 보조 인공 지능 모델에 포함된 복수의 제1 보조 레이어 각각의 출력에 기초하여 상기 제1 보조 인공 지능 모델을 재학습하고, 상기 복수의 제1 레이어 각각의 출력 및 상기 제2 보조 인공 지능 모델에 포함된 복수의 제2 보조 레이어 각각의 출력에 기초하여 상기 제2 보조 인공 지능 모델을 재학습하며, 상기 제2 학습 데이터를 상기 재학습된 제1 보조 인공 지능 모델 및 상기 재학습된 제2 보조 인공 지능 모델에 입력하고, 상기 재학습된 제1 보조 인공 지능 모델에 포함된 복수의 제1 보조 레이어 각각의 출력, 상기 재학습된 제2 보조 인공 지능 모델에 포함된 복수의 제2 보조 레이어 각각의 출력 및 상기 복수의 제2 레이어 각각의 출력에 기초하여 상기 제2 인공 지능 모델을 재학습할 수 있다. 여기서, 상기 복수의 제1 보조 레이어 각각은 대응되는 제1 레이어의 복수의 2차원 필터 각각의 크기가 축소된 복수의 필터를 포함하고, 상기 복수의 제2 보조 레이어 각각은 대응되는 제1 레이어의 복수의 2차원 필터 각각의 크기가 축소된 복수의 필터를 포함하고, 상기 복수의 제1 보조 레이어 각각에 포함된 복수의 필터 각각은 대응되는 제2 보조 레이어 각각에 포함된 필터와 형태가 상이할 수 있다.
- [20] 한편, 본 개시의 일 실시 예에 따르면, 전자 장치의 제어 방법은 제1 학습 데이터를 통해 학습된 제1 인공 지능 모델 및 상기 제1 학습 데이터를 통해 학습된 제2 인공 지능 모델 각각에 제2 학습 데이터를 입력하는 단계 및 상기 제1

인공 지능 모델에 포함된 복수의 제1 레이어 각각의 출력 및 상기 제2 인공 지능 모델에 포함된 복수의 제2 레이어 각각의 출력에 기초하여 상기 제2 인공 지능 모델을 재학습하는 단계를 포함하며, 상기 복수의 제1 레이어 각각은 복수의 2차원 필터를 포함하고, 상기 복수의 제2 레이어 각각은 대응되는 제1 레이어의 복수의 2차원 필터 각각의 크기가 축소된 복수의 필터를 포함할 수 있다.

- [21] 또한, 상기 복수의 제1 레이어의 개수는 상기 복수의 제2 레이어의 개수와 동일하고, 상기 재학습하는 단계는 상기 복수의 제1 레이어 각각의 출력을 대응되는 제2 레이어의 출력과 비교하여 복수의 비교 결과를 획득하고, 상기 복수의 비교 결과에 기초하여 상기 제2 인공 지능 모델을 재학습할 수 있다.
- [22] 그리고, 상기 복수의 제1 레이어 각각에 포함된 상기 복수의 2차원 필터의 개수는 대응되는 제2 레이어에 포함된 상기 복수의 필터의 개수와 동일하며, 상기 복수의 제1 레이어 각각의 출력의 크기는 대응되는 제2 레이어의 출력의 크기와 동일할 수 있다.
- [23] 또한, 상기 재학습하는 단계는 상기 복수의 제2 레이어 각각의 출력을 제3 인공 지능 모델 각각에 입력하고, 상기 제3 인공 지능 모델로부터 출력되는 상기 복수의 제2 레이어의 출력에 대한 복수의 구별 결과를 획득하며, 상기 복수의 비교 결과 및 상기 복수의 구별 결과에 기초하여 상기 제2 인공 지능 모델을 재학습하고, 상기 제3 인공 지능 모델은 상기 복수의 제1 레이어 각각의 출력 및 상기 복수의 제2 레이어 각각의 출력을 구별하도록 학습된 인공 지능 모델일 수 있다.
- [24] 그리고, 상기 제2 학습 데이터를 상기 재학습된 제2 인공 지능 모델 각각에 입력하는 단계 및 상기 복수의 제1 레이어 각각의 출력 및 상기 재학습된 제2 인공 지능 모델에 포함된 복수의 제3 레이어 각각의 출력을 구별하도록 상기 제3 인공 지능 모델을 재학습하는 단계를 더 포함할 수 있다.
- [25] 또한, 상기 재학습하는 단계는 상기 복수의 비교 결과 및 상기 복수의 구별 결과를 가중합하여 상기 제2 인공 지능 모델을 재학습할 수 있다.
- [26] 그리고, 상기 재학습하는 단계는 상기 복수의 제1 레이어 중 마지막 레이어에 대응되는 비교 결과의 가중치를 나머지 레이어에 대응되는 복수의 비교 결과의 가중치 이상으로 부여하여 상기 제2 인공 지능 모델을 재학습할 수 있다.
- [27] 또한, 상기 복수의 필터 각각은 $1 \times N$ 형태 또는 $N \times 1$ 형태이고, 입력 데이터를 행(row) 단위 또는 컬럼(column) 단위로 독출하는 단계 및 상기 독출된 입력 데이터를 상기 재학습된 제2 인공 지능 모델에 입력하여 상기 입력 데이터를 처리하는 단계를 더 포함할 수 있다.
- [28] 그리고, 상기 제1 인공 지능 모델 및 상기 제2 인공 지능 모델은 컨볼루션 신경망(Convolutional Neural Network, CNN) 형태로 구현될 수 있다.
- [29] 한편, 본 개시의 일 실시 예에 따르면, 전자 장치의 동작 방법을 실행하기 위한 프로그램이 저장된 비일시적 컴퓨터 판독가능 기록매체에 있어서, 상기 동작 방법은 제1 학습 데이터를 통해 학습된 제1 인공 지능 모델 및 상기 제1 학습

데이터를 통해 학습된 제2 인공 지능 모델 각각에 제2 학습 데이터를 입력하는 단계 및 상기 제1 인공 지능 모델에 포함된 복수의 제1 레이어 각각의 출력 및 상기 제2 인공 지능 모델에 포함된 복수의 제2 레이어 각각의 출력에 기초하여 상기 제2 인공 지능 모델을 재학습하는 단계를 포함하며, 상기 복수의 제1 레이어 각각은 복수의 2차원 필터를 포함하고, 상기 복수의 제2 레이어 각각은 대응되는 제1 레이어의 복수의 2차원 필터 각각의 크기가 축소된 복수의 필터를 포함할 수 있다.

발명의 효과

- [30] 이상과 같은 본 개시의 다양한 실시 예에 따르면, 전자 장치는 Teacher network 및 Student network의 각 레이어의 출력을 비교하여 Student network를 재학습함에 따라 로우-레벨(low-level)의 특징이 부각된 인공 지능 모델을 획득할 수 있다.
- [31] 또한, 전자 장치는 Teacher network 및 Student network의 각 레이어의 출력을 구분하도록 학습된 discriminator를 더 이용하여 로우-레벨의 특징이 좀더 부각된 인공 지능 모델을 획득할 수도 있다.
- [32] 그리고, 전자 장치는 Teacher network의 2차원 형태의 필터의 크기를 축소한 Student network를 이용함에 따라 인공 지능 모델의 용량을 경량화하고 on the fly 방식으로 입력 데이터를 처리할 수 있다.

도면의 간단한 설명

- [33] 도 1a 및 도 1b는 종래 기술에 따른 문제점을 설명하기 위한 도면들이다.
- [34] 도 2는 본 개시의 일 실시 예에 따른 전자 장치의 구성을 나타내는 블록도이다.
- [35] 도 3은 본 개시의 일 실시 예에 따른 제1 인공 지능 모델과 제2 인공 지능 모델을 비교하기 위한 도면이다.
- [36] 도 4는 본 개시의 일 실시 예에 따른 제1 인공 지능 모델과 제2 인공 지능 모델의 각 레이어의 출력을 비교하는 도면이다.
- [37] 도 5는 본 개시의 일 실시 예에 따른 제2 인공 지능 모델의 재학습 단계를 설명하기 위한 도면이다.
- [38] 도 6a 및 도 6b는 본 개시의 일 실시 예에 따른 on the fly 방식을 설명하기 위한 도면들이다.
- [39] 도 7a 내지 도 8b는 본 개시의 다양한 실시 예에 따른 재학습된 제2 인공 지능 모델의 성능을 설명하기 위한 도면들이다.
- [40] 도 9는 본 개시의 다른 실시 예에 따른 재학습된 제2 인공 지능 모델의 성능을 설명하기 위한 도면이다.
- [41] 도 10은 본 개시의 확장 실시 예에 따른 단계적 재학습 방법을 설명하기 위한 도면이다.
- [42] 도 11a 및 도 11b는 본 개시의 확장 실시 예에 따른 병렬적 재학습 방법을 설명하기 위한 도면들이다.
- [43] 도 12는 본 개시의 일 실시 예에 따른 전자 장치의 제어 방법을 설명하기 위한

흐름도이다.

발명의 실시를 위한 최선의 형태

[44]

-

발명의 실시를 위한 형태

[45] 이하에서는 첨부 도면을 참조하여 본 개시를 상세히 설명한다.

[46] 본 개시의 실시 예에서 사용되는 용어는 본 개시에서의 기능을 고려하면서 가능한 현재 널리 사용되는 일반적인 용어들을 선택하였으나, 이는 당 분야에 종사하는 기술자의 의도 또는 판례, 새로운 기술의 출현 등에 따라 달라질 수 있다. 또한, 특정한 경우는 출원인이 임의로 선정한 용어도 있으며, 이 경우 해당되는 개시의 설명 부분에서 상세히 그 의미를 기재할 것이다. 따라서 본 개시에서 사용되는 용어는 단순한 용어의 명칭이 아닌, 그 용어가 가지는 의미와 본 개시의 전반에 걸친 내용을 토대로 정의되어야 한다.

[47] 본 명세서에서, "가진다," "가질 수 있다," "포함한다," 또는 "포함할 수 있다" 등의 표현은 해당 특징(예: 수치, 기능, 동작, 또는 부품 등의 구성요소)의 존재를 가리키며, 추가적인 특징의 존재를 배제하지 않는다.

[48] A 또는/및 B 중 적어도 하나라는 표현은 "A" 또는 "B" 또는 "A 및 B" 중 어느 하나를 나타내는 것으로 이해되어야 한다.

[49] 본 명세서에서 사용된 "제1," "제2," "첫째," 또는 "둘째," 등의 표현들은 다양한 구성요소들을, 순서 및/또는 중요도에 상관없이 수식할 수 있고, 한 구성요소를 다른 구성요소와 구분하기 위해 사용될 뿐 해당 구성요소들을 한정하지 않는다.

[50] 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 출원에서, "포함하다" 또는 "구성되다" 등의 용어는 명세서상에 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.

[51] 본 명세서에서, 사용자라는 용어는 전자 장치를 사용하는 사람 또는 전자 장치를 사용하는 장치(예: 인공 지능 전자 장치)를 지칭할 수 있다.

[52] 이하 첨부된 도면들을 참조하여 본 개시의 일 실시 예를 보다 상세하게 설명한다.

[53] 도 2는 본 개시의 일 실시 예에 따른 전자 장치(100)의 구성을 나타내는 블록도이다.

[54] 전자 장치(100)는 인공 지능 모델을 학습하는 장치로서, TV, 데스크탑 PC, 노트북, 비디오 월(video wall), LFD(large format display), Digital Signage(디지털 간판), DID(Digital Information Display), 프로젝터 디스플레이, DVD(digital video disk) 플레이어, 스마트폰, 태블릿 PC, 모니터, 스마트 안경, 스마트 워치, 셋탑박스(STB), 스피커, 컴퓨터 본체 등과 같은 장치일 수 있다. 다만, 이에

한정되는 것은 아니며, 전자 장치(100)는 인공 지능 모델을 학습할 수 있는 장치라면 어떠한 장치라도 무방하다.

- [55] 도 2에 따르면, 전자 장치(100)는 메모리(110) 및 프로세서(120)를 포함한다. 다만, 이에 한정되는 것은 아니며, 전자 장치(100)는 일부 구성이 제외된 형태로 구현될 수도 있다.
- [56] 메모리(110)는 제1 인공 지능 모델 및 제2 인공 지능 모델을 저장할 수 있다. 예를 들어, 제1 인공 지능 모델 및 제2 인공 지능 모델은 컨볼루션 신경망(Convolutional Neural Network, CNN) 형태로 구현될 수 있다.
- [57] 제1 인공 지능 모델은 제1 학습 데이터를 통해 학습될 수 있으며, 용량이 큰 네트워크(Teacher network)일 수 있다. 그리고, 제2 인공 지능 모델은 제1 학습 데이터를 통해 학습될 수 있으며, 용량이 작은 네트워크(Student network)일 수 있다. 즉, 제1 인공 지능 모델 및 제2 인공 지능 모델은 동일한 제1 학습 데이터를 이용하여 학습될 수 있다.
- [58] 제2 인공 지능 모델은 제1 인공 지능 모델에서 필터의 크기가 축소된 인공 지능 모델일 수 있다. 예를 들어, 제1 인공 지능 모델에 포함된 복수의 제1 레이어의 개수는 제2 인공 지능 모델에 포함된 복수의 제2 레이어의 개수와 동일할 수 있다. 그리고, 복수의 제1 레이어 각각은 복수의 2차원 필터를 포함하고, 복수의 제2 레이어 각각은 대응되는 제1 레이어의 복수의 2차원 필터 각각의 크기가 축소된 복수의 필터를 포함할 수 있다. 여기서, 복수의 제1 레이어 각각에 포함된 복수의 2차원 필터의 개수는, 대응되는 제2 레이어에 포함된 복수의 필터의 개수와 동일할 수 있다.
- [59] 만약, 제1 인공 지능 모델에 포함된 모든 필터가 $N \times N$ 형태이고, 제2 인공 지능 모델에 포함된 모든 필터가 $n \times N$ 형태이면(여기서, $n < N$), 제2 인공 지능 모델은 제1 인공 지능 모델의 n/N 의 용량으로 구현될 수 있다. 다만, 상대적으로 적은 용량으로 구현되었고 동일한 학습 방법으로 학습(training)되었기 때문에, 제2 인공 지능 모델의 출력은 제1 인공 지능 모델의 출력보다 품질이 낮을 수 있다. 이후, 제2 인공 지능 모델의 출력이 제1 인공 지능 모델의 출력과 유사해지도록 제2 인공 지능 모델의 재학습이 수행되며, 이에 대하여는 프로세서(120)를 동작에서 설명한다.
- [60] 한편, 메모리(110)는 제1 학습 데이터를 통해 학습된 제1 인공 지능 모델 및 제2 인공 지능 모델을 저장하는 것으로 설명하였으나, 이에 한정되는 것은 아니다. 예를 들어, 메모리(110)는 제1 학습 데이터, 웨이트 값이 결정되지 않은 제1 인공 지능 모델 및 웨이트 값이 결정되지 않은 제2 인공 지능 모델만을 저장할 수도 있다. 가령, 웨이트 값이 결정되지 않은 제1 인공 지능 모델은 학습이 수행되기 전의 모델로서, 제1 인공 지능 모델과 형태가 동일하나 모든 웨이트 값이 0일 수 있다. 웨이트 값이 결정되지 않은 제2 인공 지능 모델은 학습이 수행되기 전의 모델로서, 제2 인공 지능 모델과 형태가 동일하나 모든 웨이트 값이 0일 수 있다. 이 경우, 프로세서(120)는 제1 학습 데이터를 이용하여 웨이트 값이 결정되지

않은 제1 인공 지능 모델 및 웨이트 값이 결정되지 않은 제2 인공 지능 모델 각각을 학습하여 제1 인공 지능 모델 및 제2 인공 지능 모델을 획득하고, 획득된 제1 인공 지능 모델 및 제2 인공 지능 모델을 메모리(110)에 저장할 수도 있다.

[61] 메모리(110)는 복수의 제1 레이어 각각의 출력 및 복수의 제2 레이어 각각의 출력을 구별하도록 학습된 제3 인공 지능 모델에 대한 정보를 더 저장할 수 있다. 제3 인공 지능 모델은 discriminator의 일종으로, 데이터가 입력되면 해당 데이터가 제1 인공 지능 모델의 출력인지 또는 제2 인공 지능 모델의 출력인지를 식별하도록 학습된 모델일 수 있다.

[62] 다만, 이에 한정되는 것은 아니면, 메모리(110)는 웨이트 값이 결정되지 않은 제3 인공 지능 모델을 저장할 수도 있다. 가령, 웨이트 값이 결정되지 않은 제3 인공 지능 모델은 학습이 수행되기 전의 모델로서, 제3 인공 지능 모델과 형태가 동일하나 모든 웨이트 값이 0일 수 있다. 이 경우, 프로세서(120)는 복수의 제1 레이어 각각의 출력 및 복수의 제2 레이어 각각의 출력을 구별하도록 웨이트 값이 결정되지 않은 제3 인공 지능 모델을 학습시키고, 학습이 완료된 제3 인공 지능 모델을 메모리(110)에 저장할 수도 있다.

[63] 메모리(110)는 제2 학습 데이터를 더 저장할 수 있다. 제2 학습 데이터는 제2 인공 지능 모델의 재학습을 수행하는 경우에 이용되는 데이터일 수 있다.

[64] 메모리(110)는 입력 데이터를 더 저장할 수 있다. 입력 데이터는 재학습된 제2 인공 지능 모델을 통해 처리되는 대상일 수 있다.

[65] 프로세서(120)는 전자 장치(100)의 동작을 전반적으로 제어한다. 구체적으로, 프로세서(120)는 전자 장치(100)의 각 구성과 연결되어 전자 장치(100)의 동작을 전반적으로 제어할 수 있다. 예를 들어, 프로세서(120)는 메모리(110), 통신 인터페이스(미도시) 등과 같은 구성과 연결되어 전자 장치(100)의 동작을 제어할 수 있다.

[66] 일 실시 예에 따라 프로세서(120)는 디지털 시그널 프로세서(digital signal processor(DSP), 마이크로 프로세서(microprocessor), TCON(Time controller)으로 구현될 수 있다. 다만, 이에 한정되는 것은 아니며, 중앙처리장치(central processing unit(CPU)), MCU(Micro Controller Unit), MPU(micro processing unit), 컨트롤러(controller), 어플리케이션 프로세서(application processor(AP)), 또는 커뮤니케이션 프로세서(communication processor(CP)), ARM 프로세서 중 하나 또는 그 이상을 포함하거나, 해당 용어로 정의될 수 있다. 또한, 프로세서(120)는 프로세싱 알고리즘이 내장된 SoC(System on Chip), LSI(large scale integration)로 구현될 수도 있고, FPGA(Field Programmable gate array) 형태로 구현될 수도 있다.

[67] 프로세서(120)는 제2 학습 데이터를 제1 인공 지능 모델 및 제2 인공 지능 모델 각각에 입력하고, 제1 인공 지능 모델에 포함된 복수의 제1 레이어 각각의 출력 및 제2 인공 지능 모델에 포함된 복수의 제2 레이어 각각의 출력에 기초하여 제2 인공 지능 모델을 재학습할 수 있다. 여기서, 제1 인공 지능 모델에 포함된 복수의 제1 레이어 각각의 출력 및 제2 인공 지능 모델에 포함된 복수의 제2

레이어 각각의 출력은 피쳐 맵(feature map) 데이터 및 최종 출력 데이터를 포함할 수 있다. 즉, 제1 인공 지능 모델에 포함된 복수의 제1 레이어 각각의 출력 및 제2 인공 지능 모델에 포함된 복수의 제2 레이어 각각의 출력은 컨볼루션 레이어 출력(convolution layer output)(feature map) 및 FC 레이어 출력(fully connected layer output)(Logits)을 포함할 수 있다.

- [68] 프로세서(120)는 복수의 제1 레이어 각각의 출력을 대응되는 제2 레이어의 출력과 비교하여 복수의 비교 결과를 획득하고, 복수의 비교 결과에 기초하여 제2 인공 지능 모델을 재학습할 수 있다. 예를 들어, 프로세서(120)는 복수의 제1 레이어 각각의 출력과 대응되는 제2 레이어의 출력 간의 유클리디안 거리(euclidean distance)를 사용한 norm 값을 줄이도록 제2 인공 지능 모델을 재학습할 수 있다. 이러한 학습을 통해 복수의 제2 레이어 각각의 출력이 대응되는 제1 레이어의 출력과 유사해질 수 있다.
- [69] 여기서, 복수의 제1 레이어 각각의 출력의 크기는 대응되는 제2 레이어의 출력의 크기와 동일할 수 있다. 이는, 제2 인공 지능 모델이 제1 인공 지능 모델에서 필터의 크기만이 축소된 인공 지능 모델이기 때문이다.
- [70] 또는, 프로세서(120)는 복수의 제2 레이어 각각의 출력을 제3 인공 지능 모델 각각에 입력하고, 제3 인공 지능 모델로부터 출력되는 복수의 제2 레이어의 출력에 대한 복수의 구별 결과를 획득하며, 복수의 비교 결과 및 복수의 구별 결과에 기초하여 제2 인공 지능 모델을 재학습할 수도 있다. 즉, 프로세서(120)는 복수의 제2 레이어 각각의 출력이 제3 인공 지능 모델에 의해 제1 인공 지능 모델의 출력으로서 식별되도록, 제2 인공 지능 모델을 재학습할 수 있다.
- [71] 프로세서(120)는 복수의 비교 결과 및 복수의 구별 결과를 가중합하여 제2 인공 지능 모델을 재학습할 수 있다. 예를 들어, 프로세서(120)는 복수의 제1 레이어 중 마지막 레이어에 대응되는 비교 결과의 가중치를 나머지 레이어에 대응되는 복수의 비교 결과의 가중치 이상으로 부여하여 제2 인공 지능 모델을 재학습할 수도 있다.
- [72] 다만, 이에 한정되는 것은 아니며, 프로세서(120)는 얼마든지 다양한 방식으로 가중치를 부여할 수도 있다.
- [73] 이상과 같은 방법을 통해 제2 인공 지능 모델을 재학습함으로써, 제2 인공 지능 모델의 출력이 좀더 제1 인공 지능 모델의 출력과 유사해질 수 있다.
- [74] 한편, 프로세서(120)는 제2 학습 데이터를 재학습된 제2 인공 지능 모델 각각에 입력하고, 복수의 제1 레이어 각각의 출력 및 재학습된 제2 인공 지능 모델에 포함된 복수의 제3 레이어 각각의 출력을 구별하도록 제3 인공 지능 모델을 재학습할 수도 있다.
- [75] 그리고, 프로세서(120)는 제3 학습 데이터를 제1 인공 지능 모델 및 재학습된 제2 인공 지능 모델 각각에 입력하고, 제1 인공 지능 모델에 포함된 복수의 제1 레이어 각각의 출력 및 재학습된 제2 인공 지능 모델에 포함된 복수의 제3 레이어 각각의 출력에 기초하여 재학습된 제2 인공 지능 모델을 2차 재학습할

수도 있다. 가령, 프로세서(120)는 복수의 제1 레이어 각각의 출력을 대응되는 제3 레이어의 출력과 비교하여 복수의 비교 결과를 획득하고, 복수의 비교 결과에 기초하여 재학습된 제2 인공 지능 모델을 2차 재학습할 수도 있다.

[76] 이때, 프로세서(120)는 복수의 제3 레이어 각각의 출력을 재학습된 제3 인공 지능 모델 각각에 입력하고, 재학습된 제3 인공 지능 모델로부터 출력되는 복수의 제3 레이어의 출력에 대한 복수의 구별 결과를 획득하며, 복수의 비교 결과 및 복수의 구별 결과에 기초하여 재학습된 제2 인공 지능 모델을 2차 재학습할 수도 있다.

[77] 즉, 프로세서(120)는 제2 인공 지능 모델 및 제3 인공 지능 모델을 반복적으로 재학습하여 제2 인공 지능 모델의 성능을 개선할 수도 있다.

[78] 한편, 복수의 제2 레이어 각각에 포함된 복수의 필터는 대응되는 제1 레이어의 복수의 2차원 필터 각각의 크기가 축소된 형태일 수 있다. 예를 들어, 복수의 2차원 필터 각각의 크기가 $N \times N$ 형태인 경우, 복수의 필터는 $n \times N$ 형태 또는 $N \times n$ 형태일 수 있다. 여기서, n 은 N 보다 작다.

[79] 특히, 복수의 제2 레이어 각각에 포함된 복수의 필터는 대응되는 제1 레이어의 복수의 2차원 필터 각각의 크기가 축소된 1차원 형태일 수도 있다. 이 경우, 복수의 1차원 필터 각각은 $1 \times N$ 형태 또는 $N \times 1$ 형태이고, 프로세서(120)는 메모리(110)로부터 입력 데이터를 행(row) 단위 또는 컬럼(column) 단위로 독출하고, 독출된 입력 데이터를 재학습된 제2 인공 지능 모델에 입력하여 입력 데이터를 처리할 수 있다.

[80] 즉, 이러한 필터의 형태 상의 특징으로 인해, 재학습된 제2 인공 지능 모델을 이용하는 경우, on the fly 방식으로 입력 데이터를 처리할 수 있다.

[81] 한편, 이상에서는 제1 인공 지능 모델을 압축시켜 제2 인공 지능 모델을 획득하는 것으로 설명하였으나, 이러한 방법은 단계적 또는 병렬적으로 수행될 수도 있다.

[82] 예를 들어, 메모리(110)는 제1 학습 데이터를 통해 학습된 보조 인공 지능 모델에 대한 정보를 더 저장할 수 있다. 프로세서(120)는 제2 학습 데이터를 보조 인공 지능 모델에 입력하고, 복수의 제1 레이어 각각의 출력 및 보조 인공 지능 모델에 포함된 복수의 보조 레이어 각각의 출력에 기초하여 보조 인공 지능 모델을 재학습하며, 제2 학습 데이터를 재학습된 보조 인공 지능 모델에 입력하고, 재학습된 보조 인공 지능 모델에 포함된 복수의 보조 레이어 각각의 출력 및 복수의 제2 레이어 각각의 출력에 기초하여 제2 인공 지능 모델을 재학습할 수 있다. 여기서, 복수의 보조 레이어 각각은 대응되는 제1 레이어의 복수의 2차원 필터 각각의 크기가 축소된 복수의 필터를 포함할 수 있다.

[83] 즉, 보조 인공 지능 모델 역시 제1 인공 지능 모델 및 제2 인공 지능 모델과 그 형태가 거의 동일하다. 다만, 보조 인공 지능 모델은 제1 인공 지능 모델에서 필터의 크기가 축소된 인공 지능 모델일 수 있다. 또한, 단계적인 재학습의 이점을 살리기 위해서, 보조 인공 지능 모델은 제2 인공 지능 모델의 필터보다

크기가 클 수 있다.

- [84] 또는, 메모리(110)는 제1 학습 데이터를 통해 학습된 제1 보조 인공 지능 모델에 대한 정보 및 제1 학습 데이터를 통해 학습된 제2 보조 인공 지능 모델에 대한 정보를 더 저장할 수도 있다. 프로세서(120)는 제2 학습 데이터를 제1 보조 인공 지능 모델 및 제2 보조 인공 지능 모델에 입력하고, 복수의 제1 레이어 각각의 출력 및 제1 보조 인공 지능 모델에 포함된 복수의 제1 보조 레이어 각각의 출력에 기초하여 제1 보조 인공 지능 모델을 재학습하고, 복수의 제1 레이어 각각의 출력 및 제2 보조 인공 지능 모델에 포함된 복수의 제2 보조 레이어 각각의 출력에 기초하여 제2 보조 인공 지능 모델을 재학습하며, 제2 학습 데이터를 재학습된 제1 보조 인공 지능 모델 및 재학습된 제2 보조 인공 지능 모델에 입력하고, 재학습된 제1 보조 인공 지능 모델에 포함된 복수의 제1 보조 레이어 각각의 출력, 재학습된 제2 보조 인공 지능 모델에 포함된 복수의 제2 보조 레이어 각각의 출력 및 복수의 제2 레이어 각각의 출력에 기초하여 제2 인공 지능 모델을 재학습할 수 있다. 여기서, 복수의 제1 보조 레이어 각각은 대응되는 제1 레이어의 복수의 2차원 필터 각각의 크기가 축소된 복수의 필터를 포함하고, 복수의 제2 보조 레이어 각각은 대응되는 제1 레이어의 복수의 2차원 필터 각각의 크기가 축소된 복수의 필터를 포함하고, 복수의 제1 보조 레이어 각각에 포함된 복수의 필터 각각은 대응되는 제2 보조 레이어 각각에 포함된 필터와 형태가 상이할 수 있다.
- [85] 예를 들어, 프로세서(120)는 재학습된 제1 보조 인공 지능 모델에 포함된 복수의 제1 보조 레이어 각각의 출력 및 재학습된 제2 보조 인공 지능 모델에 포함된 복수의 제2 보조 레이어 각각의 출력을 가중합하고, 가중합된 결과 및 복수의 제2 레이어 각각의 출력에 기초하여 제2 인공 지능 모델을 재학습할 수 있다.
- [86] 즉, 프로세서(120)는 두 개의 보조 인공 지능 모델을 병렬적으로 재학습하고, 두 개의 재학습된 보조 인공 지능 모델에 기초하여 제2 인공 지능 모델을 재학습할 수도 있다.
- [87] 한편, 본 개시에 따른 인공 지능과 관련된 기능은 프로세서(120)와 메모리(110)를 통해 동작된다.
- [88] 프로세서(120)는 하나 또는 복수의 프로세서로 구성될 수 있다. 이때, 하나 또는 복수의 프로세서는 CPU, AP, DSP 등과 같은 범용 프로세서, GPU, VPU(Vision Processing Unit)와 같은 그래픽 전용 프로세서 또는 NPU와 같은 인공 지능 전용 프로세서일 수 있다.
- [89] 하나 또는 복수의 프로세서는, 메모리(110)에 저장된 기 정의된 동작 규칙 또는 인공 지능 모델에 따라, 입력 데이터를 처리하도록 제어한다. 또는, 하나 또는 복수의 프로세서가 인공 지능 전용 프로세서인 경우, 인공 지능 전용 프로세서는 특정 인공 지능 모델의 처리에 특화된 하드웨어 구조로 설계될 수 있다. 기 정의된 동작 규칙 또는 인공 지능 모델은 학습을 통해 만들어진 것을 특징으로

한다.

- [90] 여기서, 학습을 통해 만들어진다는 것은, 기본 인공지능 모델이 학습 알고리즘에 의하여 다수의 학습 데이터들을 이용하여 학습됨으로써, 원하는 특성(또는, 목적)을 수행하도록 설정된 기 정의된 동작 규칙 또는 인공지능 모델이 만들어짐을 의미한다. 이러한 학습은 본 개시에 따른 인공지능이 수행되는 기기 자체에서 이루어질 수도 있고, 별도의 서버 및/또는 시스템을 통해 이루어 질 수도 있다. 학습 알고리즘의 예로는, 지도형 학습(supervised learning), 비지도형 학습(unsupervised learning), 준지도형 학습(semi-supervised learning) 또는 강화 학습(reinforcement learning)이 있으나, 전술한 예에 한정되지 않는다.
- [91] 인공지능 모델은 복수의 신경망 레이어들로 구성될 수 있다. 복수의 신경망 레이어들 각각은 복수의 가중치들(weight values)을 갖고 있으며, 이전(previous) 레이어의 연산 결과와 복수의 가중치들 간의 연산을 통해 신경망 연산을 수행한다. 복수의 신경망 레이어들이 갖고 있는 복수의 가중치들은 인공지능 모델의 학습 결과에 의해 최적화될 수 있다. 예를 들어, 학습 과정 동안 인공지능 모델에서 획득한ロス(loss) 값 또는 코스트(cost) 값이 감소 또는 최소화되도록 복수의 가중치들이 갱신될 수 있다.
- [92] 인공지능 신경망은 심층 신경망(DNN:Deep Neural Network)를 포함할 수 있으며, 예를 들어, CNN(Convolutional Neural Network), DNN(Deep Neural Network), RNN(Recurrent Neural Network), RBM(Restricted Boltzmann Machine), DBN(Deep Belief Network), BRDNN(Bidirectional Recurrent Deep Neural Network), GAN(Generative Adversarial Network) 또는 심층 Q-네트워크(Deep Q-Networks) 등이 있으나, 이에 한정되는 것은 아니다.
- [93] 한편, 전자 장치(100)는 통신 인터페이스를 더 포함할 수도 있다.
- [94] 통신 인터페이스는 다양한 유형의 통신방식에 따라 다양한 유형의 외부 장치와 통신을 수행하는 구성이다. 통신 인터페이스는 와이파이 모듈, 블루투스 모듈, 적외선 통신 모듈 및 무선 통신 모듈 등을 포함한다. 여기서, 각 통신 모듈은 적어도 하나의 하드웨어칩 형태로 구현될 수 있다.
- [95] 프로세서(120)는 통신 인터페이스를 이용하여 각종 외부 장치와 통신을 수행할 수 있다. 예를 들어, 프로세서(120)는 통신 인터페이스를 통해 제1 외부 장치로부터 제1 학습 데이터, 제2 학습 데이터, 제1 인공지능 모델, 제2 인공지능 모델 및 제3 인공지능 모델을 수신하고, 수신된 데이터를 메모리(110)에 저장할 수 있다. 그리고, 프로세서(120)는 재학습된 제2 인공지능 모델을 제2 외부 장치로 전송하도록 통신 인터페이스를 제어할 수도 있다. 여기서, 외부 장치는 서버, 블루투스 이어폰, 디스플레이 장치 등을 포함할 수 있다.
- [96] 와이파이 모듈, 블루투스 모듈은 각각 WiFi 방식, 블루투스 방식으로 통신을 수행한다. 와이파이 모듈이나 블루투스 모듈을 이용하는 경우에는 SSID 및 세션 키 등과 같은 각종 연결 정보를 먼저 송수신하여, 이를 이용하여 통신 연결한 후

- 각종 정보들을 송수신할 수 있다.
- [97] 적외선 통신 모듈은 시 광선과 밀리미터파 사이에 있는 적외선을 이용하여 근거리에서 무선으로 데이터를 전송하는 적외선 통신(IrDA, infrared Data Association)기술에 따라 통신을 수행한다.
- [98] 무선 통신 모듈은 상술한 통신 방식 이외에 지그비(zigbee), 3G(3rd Generation), 3GPP(3rd Generation Partnership Project), LTE(Long Term Evolution), LTE-A(LTE Advanced), 4G(4th Generation), 5G(5th Generation)등과 같은 다양한 무선 통신 규격에 따라 통신을 수행하는 적어도 하나의 통신 칩을 포함할 수 있다.
- [99] 그 밖에 통신 인터페이스는 LAN(Local Area Network) 모듈, 이더넷 모듈, 또는 패어 케이블, 동축 케이블 또는 광섬유 케이블 등을 이용하여 통신을 수행하는 유선 통신 모듈 중 적어도 하나를 포함할 수 있다.
- [100] 통신 인터페이스는 입출력 인터페이스를 더 포함할 수 있다. 입출력 인터페이스는 HDMI(High Definition Multimedia Interface), MHL (Mobile High-Definition Link), USB (Universal Serial Bus), DP(Display Port), 썬더볼트(Thunderbolt), VGA(Video Graphics Array)포트, RGB 포트, D-SUB(D-subminiature), DVI(Digital Visual Interface) 중 어느 하나의 인터페이스일 수 있다.
- [101] 입출력 인터페이스는 오디오 및 비디오 신호 중 적어도 하나를 입출력 할 수 있다.
- [102] 구현 예에 따라, 입출력 인터페이스는 오디오 신호만을 입출력하는 포트와 비디오 신호만을 입출력하는 포트를 별개의 포트에 포함하거나, 오디오 신호 및 비디오 신호를 모두 입출력하는 하나의 포트에 구현될 수 있다.
- [103] 이상과 같은 본 개시의 다양한 실시 예에 따르면, 제2 인공 지능 모델이 재학습됨에 따라 성능을 좀더 개선하면서도 제1 인공 지능 모델의 경량화가 가능하고, 제2 인공 지능 모델에 포함된 필터의 형태 상의 특징으로 인해 on the fly 방식으로 입력 데이터를 처리할 수 있다.
- [104] 이하에서는 다양한 도면을 통해 프로세서(120)의 동작을 좀더 구체적으로 설명한다.
- [105] 도 3은 본 개시의 일 실시 예에 따른 제1 인공 지능 모델과 제2 인공 지능 모델을 비교하기 위한 도면이다.
- [106] 제1 인공 지능 모델(Teacher Configuration) 및 제2 인공 지능 모델(Student Configuration)은 그 형태가 거의 동일하다. 다만, 제2 인공 지능 모델은 제1 인공 지능 모델에서 필터의 크기가 축소된 인공 지능 모델일 수 있다. 예를 들어, 도 3에 도시된 바와 같이, 제1 인공 지능 모델에 포함된 복수의 제1 레이어의 개수 및 제2 인공 지능 모델에 포함된 복수의 제2 레이어의 개수는 총 5개로 동일할 수 있다. 그리고, 대응되는 레이어의 채널의 개수, 즉 대응되는 레이어에 포함된 필터의 개수도 동일할 수 있다. 가령, 제1 인공 지능 모델의 첫 번째 레이어 및 제2 인공 지능 모델의 첫 번째 레이어는 총 64 채널을 포함할 수 있다.

- [107] 그리고, 복수의 제1 레이어 각각은 복수의 2차원 필터를 포함하고, 복수의 제2 레이어 각각은 대응되는 제1 레이어의 복수의 2차원 필터 각각의 크기가 축소된 복수의 필터를 포함할 수 있다. 도 3에서는 설명의 편의를 위해 복수의 제1 레이어 각각이 복수의 3×3 형태의 필터를 포함하고, 복수의 제2 레이어 각각이 복수의 1×3 형태의 필터를 포함하는 것으로 도시하였다. 따라서, 제2 인공 지능 모델은 제1 인공 지능 모델과 비교하여 1/3의 용량으로 구현될 수 있다.
- [108] 또한, 제1 인공 지능 모델에 포함된 복수의 제1 레이어의 개수 및 제2 인공 지능 모델에 포함된 복수의 제2 레이어의 개수가 동일하여, 1:1로 feature map의 비교가 가능하다.
- [109] 그리고, 필터의 형태에 차이가 있으나, 컨볼루션 연산 결과는 동일한 형태로 출력될 수 있다. 즉, 복수의 제1 레이어 각각의 출력의 크기는 대응되는 제2 레이어의 출력의 크기와 동일할 수 있으며, 그에 따라 대응되는 출력간 직접적인 비교가 가능하며, 별도로 크기 변경 등의 동작을 수행할 필요가 없다.
- [110] 도 4는 본 개시의 일 실시 예에 따른 제1 인공 지능 모델과 제2 인공 지능 모델의 각 레이어의 출력을 비교하는 도면이다.
- [111] 프로세서(120)는 제1 인공 지능 모델에 포함된 복수의 제1 레이어 각각의 출력을 대응되는 제2 인공 지능 모델에 포함된 제2 레이어의 출력과 비교하여 복수의 비교 결과를 획득하고, 복수의 제2 레이어 각각의 출력을 제3 인공 지능 모델 각각에 입력하여 제3 인공 지능 모델로부터 출력되는 복수의 제2 레이어의 출력에 대한 복수의 구별 결과를 획득할 수 있다.
- [112] 즉, 프로세서(120)는 복수의 제1 레이어 중 첫 번째 레이어의 출력과 복수의 제2 레이어 중 첫 번째 레이어의 출력을 비교하여 첫 번째 레이어 간의 비교 결과를 획득하고, 복수의 제2 레이어 중 첫 번째 레이어의 출력에 대한 제3 인공 지능 모델의 구별 결과를 획득할 수 있다. 도 4에서는 각 레이어에서의 비교 결과(L2 loss, Softmax) 및 구별 결과(Discriminator loss)를 Guide Loss로서 표시하였으며, 프로세서(120)는 첫 번째 레이어 이후의 레이어에 대하여도 동일한 Guide Loss를 획득할 수 있다.
- [113] 프로세서(120)는 각 레이어의 Guide Loss를 반영하여 제2 인공 지능 모델을 재학습함에 따라 로우-레벨의 특징을 좀더 부각시킬 수 있다. 즉, 마지막 softmax 정보만을 비교하는 종래와 비교하여, 본 개시에 의하면 네트워크의 중간 결과가 가이드되어 제2 인공 지능 모델의 성능이 개선될 수 있다.
- [114] 도 5는 본 개시의 일 실시 예에 따른 제2 인공 지능 모델의 재학습 단계를 설명하기 위한 도면이다.
- [115] 먼저, 제1 인공 지능 모델(Teacher) 및 제2 인공 지능 모델(Student)은 제1 학습 데이터를 통해 초기 학습이 수행된 상태일 수 있다. 그리고, 제3 인공 지능 모델(Discriminator)은 제1 인공 지능 모델의 피쳐 맵 데이터 및 제2 인공 지능 모델의 피쳐 맵 데이터를 구별하도록 학습된 상태일 수 있다.
- [116] 프로세서(120)는 제2 학습 데이터(Xi)를 제1 인공 지능 모델 및 제2 인공 지능

모델 각각에 입력하고, 복수의 제1 레이어 각각의 출력을 대응되는 제2 레이어의 출력과 비교하여 복수의 비교 결과(L2 loss, Softmax)를 획득할 수 있다.

[117] 그리고, 프로세서(120)는 복수의 제2 레이어 각각의 출력을 제3 인공 지능 모델 각각에 입력하여 복수의 제2 레이어의 출력에 대한 복수의 구별 결과(Discriminator loss)를 획득할 수 있다.

[118] 프로세서(120)는 복수의 비교 결과 및 복수의 구별 결과에 기초하여 제2 인공 지능 모델을 재학습할 수 있다.

[119] 또한, 프로세서(120)는 제2 학습 데이터를 재학습된 제2 인공 지능 모델 각각에 입력하고, 복수의 제1 레이어 각각의 출력 및 재학습된 제2 인공 지능 모델에 포함된 복수의 제3 레이어 각각의 출력을 구별하도록 제3 인공 지능 모델을 재학습할 수 있다.

[120] 이상의 과정을 반복함에 따라, 제2 인공 지능 모델의 성능이 개선되고, 제3 인공 지능 모델의 구별 성능이 개선될 수 있다.

[121] 도 6a 및 도 6b는 본 개시의 일 실시 예에 따른 on the fly 방식을 설명하기 위한 도면들이다.

[122] 도 6a는 종래의 2차원 필터를 이용하는 경우를 설명하기 위한 도면으로, 3×3 의 2차원 필터를 가정하였다. 2차원 필터는 입력 데이터의 제1 영역(610)부터 우측으로 시프트되며, 컨볼루션 연산이 수행될 수 있다. 이 경우, 프로세서는 메모리로부터 제1 영역(610)을 독출하여 연산을 수행한 후, 제1 영역(610)으로부터 한 칸 우측으로 시프트된 영역을 연산하게 된다. 이때, 프로세서는 추가되는 영역만을 메모리로부터 독출하고, 제1 영역(610) 중 중첩되지 않는 영역을 삭제할 수 있다. 이러한 과정이 반복되어 행이 변경된 후, 프로세서는 제2 영역(620)을 독출하게 된다. 즉, 제1 영역(610) 중 일부가 메모리로부터 다시 독출될 수 있다. 이 경우, 제3 영역(630)은 총 3회 독출될 수 있다.

[123] 이에 대해, 도 6b는 1차원 필터를 이용하는 경우를 설명하기 위한 도면으로, 1×3 의 1차원 필터를 가정하였다. 1차원 필터는 입력 데이터의 제4 영역(640)부터 우측으로 시프트되며, 컨볼루션 연산이 수행될 수 있다. 이 경우, 프로세서는 메모리(110)로부터 제4 영역(640)을 독출하여 연산을 수행한 후, 제4 영역(640)으로부터 한 칸 우측으로 시프트된 영역을 연산하게 된다. 이후, 행이 변경된 후, 프로세서(120)는 제5 영역(650)을 독출하게 되나 제4 영역(640)과 중첩되는 부분이 없기 때문에 동일한 영역을 반복 독출하는 문제가 해결되며, on the fly 방식에 좀더 유리한 연산이 가능하다.

[124] 도 7a 내지 도 8b는 본 개시의 다양한 실시 예에 따른 재학습된 제2 인공 지능 모델의 성능을 설명하기 위한 도면들이다.

[125] 도 7a 내지 도 8b는 도 3과 같은 형태의 제1 인공 지능 모델 및 제2 인공 지능 모델을 이용한 실험 결과로서, 제1 인공 지능 모델 및 제2 인공 지능 모델은 각각 8개의 컨볼루션 레이어(Convolution layer), 1개의 FC 레이어(Fully connected

layer)를 포함하는 것으로 가정하였다. 그리고, 제1 인공 지능 모델의 복수의 제1 레이어 각각은 복수의 3×3 형태의 필터를 포함하는 것으로 가정하였다. 다만, 제2 인공 지능 모델은 도 7a 및 도 7b의 x축에 도시된 바와 같이, 복수의 제2 레이어 중 일부 레이어만을 1×3 형태의 필터로 변경하였다. 도 8a 및 도 8b는 제2 인공 지능 모델의 모든 레이어를 1×3 형태의 필터로 변경하였다.

- [126] 도 7a 내지 도 8b에서 Teacher는 제1 인공 지능 모델의 Classification Accuracy 성능을 나타내고, Student는 재학습된 제2 인공 지능 모델의 Classification Accuracy 성능을 나타내며, Baseline은 단순히 1×3 형태만을 이용한 모델의 Classification Accuracy 성능을 나타낸다.
- [127] 도 7a 및 도 7b는 컨볼루션 레이어의 출력과 FC 레이어의 출력만을 고려하여 재학습된 제2 인공 지능 모델의 Classification Accuracy 성능을 나타내며, Student가 Baseline보다 개선된 성능을 나타낸다.
- [128] 다만, 첫 번째 컨볼루션 레이어는 분류나 객체 검출에 중요하다고 여겨지는 Edge, Curve 등과 같은 로우-레벨(low-level)의 특징을 추출하는데 매우 중요한 역할을 수행하고, input 사이즈도 가장 크다. 따라서, 첫 번째 컨볼루션 레이어까지 1×3 형태의 필터로 바꾸는 경우 급격한 성능 저하가 발생할 수 있다.
- [129] 이러한 문제는 제3 인공 지능 모델을 이용하여 해결이 가능하다. 도 8a 및 도 8b는 컨볼루션 레이어의 출력과 FC 레이어의 출력 및 제3 인공 지능 모델의 구별 결과를 고려하여 재학습된 제2 인공 지능 모델의 Classification Accuracy 성능을 나타낸다. 다만, 도 8a 및 도 8b의 x축에 도시된 바와 같이, 제3 인공 지능 모델은 일부 레이어에만 적용되었다. 이 경우, Student가 Baseline보다 개선된 성능을 나타낼 수 있다.
- [130] 도 9는 본 개시의 일 실시 예에 따른 종래 기술과 재학습된 제2 인공 지능 모델의 성능을 비교하기 위한 도면이다.
- [131] Baseline은 1×3 형태만을 이용한 모델의 성능을 나타내며, L_{HT} 는 종래 기술의 성능을 나타낸다. $L_{HT} + L_{L2}$ 는 종래 기술과 본 개시의 각 레이어의 비교 결과를 이용한 경우의 성능을 나타내고, $L_{HT} + L_{Gen}$ 은 종래 기술과 본 개시의 각 레이어의 구별 결과를 이용한 경우의 성능을 나타내며, $L_{HT} + L_{L2} + L_{Gen}$ 은 종래 기술과 본 개시의 각 레이어의 비교 결과 및 구별 결과를 이용한 경우의 성능을 나타내며, 비교 결과 및 구별 결과를 모두 이용한 경우의 성능이 가장 개선되었음을 알 수 있다.
- [132] 도 10은 본 개시의 확장 실시 예에 따른 단계적 재학습 방법을 설명하기 위한 도면이다.
- [133] 프로세서(120)는 도 10에 도시된 바와 같이, 복수의 보조 인공 지능 모델(TA)를 더 이용하여 최종적으로 제2 인공 지능 모델을 재학습할 수도 있다. 여기서, 보조 인공 지능 모델을 재학습하는 방법은 이상에서 제2 인공 지능 모델을 재학습하는 방법과 동일하므로 중복되는 설명은 생략한다.

- [134] 다만, 보조 인공지능 모델에 포함된 필터는 제1 인공지능 모델에 포함된 필터보다 크기가 축소될 수 있다. 반면, 보조 인공지능 모델에 포함된 필터는 제2 인공지능 모델에 포함된 필터보다 크기가 클 수 있다.
- [135] 즉, 제1 인공지능 모델을 제2 인공지능 모델로 한번에 크기를 축소시키는 경우, 중요 인자의 손실이 있을 수 있으나, 제1 인공지능 모델을 적어도 하나의 보조 인공지능 모델을 이용하여 단계적으로 크기를 축소시킴에 따라 중요 인자의 보존이 가능하다. 즉, 단계적으로 재학습된 제2 인공지능 모델은 보조 인공지능 모델을 이용하지 않고 재학습된 제2 인공지능 모델보다 제1 인공지능 모델의 특성을 구비할 수 있다.
- [136] 도 11a 및 도 11b는 본 개시의 확장 실시 예에 따른 병렬적 재학습 방법을 설명하기 위한 도면들이다.
- [137] 프로세서(120)는 도 11a에 도시된 바와 같이, 복수의 보조 인공지능 모델(TA)을 병렬적으로 이용할 수도 있다. 특히, 프로세서(120)는 필터의 형태가 상이한 복수의 보조 인공지능 모델을 이용함에 따라 제1 인공지능 모델의 특성을 좀더 반영한 제2 인공지능 모델을 획득할 수도 있다.
- [138] 예를 들어, 프로세서(120)는 도 11b에 도시된 바와 같이, 1×1 형태의 필터를 포함하는 제1 보조 인공지능 모델, 1×3 형태의 필터를 포함하는 제2 보조 인공지능 모델, 3×1 형태의 필터를 포함하는 제3 보조 인공지능 모델 및 2×2 형태의 필터를 포함하는 제4 보조 인공지능 모델을 각각 재학습할 수 있다. 제1 내지 제4 보조 인공지능 모델을 재학습하는 방법은 이상에서 제2 인공지능 모델을 재학습하는 방법을 병렬적으로 적용한 것에 불과하여 중복되는 설명은 생략한다.
- [139] 이 경우, 제1 인공지능 모델의 특성에 적합한 필터의 형태를 갖는 보조 인공지능 모델의 웨이트 값은 다른 보조 인공지능 모델의 웨이트 값보다 클 수 있다. 가령, 제1 인공지능 모델의 특성을 제4 보조 인공지능 모델의 2×2 형태의 필터가 가장 잘 반영할 수 있다면, 제4 보조 인공지능 모델의 웨이트 값은 나머지 보조 인공지능 모델의 웨이트 값보다 클 수 있다.
- [140] 반면, 보조 인공지능 모델을 이용하지 않는다면, 제2 인공지능 모델의 1×3 형태의 필터는 제1 인공지능 모델의 특성을 상대적으로 덜 반영하게 된다. 따라서, 복수의 보조 인공지능 모델을 이용하게 되면, 도 10과 같은 단계적 재학습의 효과도 동일하게 유지하면서 제1 인공지능 모델의 특성을 좀더 잘 반영할 수 있다.
- [141] 도 11b에서는 재학습된 제1 내지 제4 보조 인공지능 모델을 평균하여 제2 인공지능 모델을 재학습하는 것으로 도시하였으나, 이에 한정되는 것은 아니다. 예를 들어, 프로세서(120)는 재학습된 제1 내지 제4 보조 인공지능 모델을 가중합하여 제2 인공지능 모델을 재학습할 수도 있다.
- [142] 또한, 프로세서(120)는 얼마든지 다양한 개수 또는 다양한 형태의 보조 인공지능 모델을 이용할 수도 있다.

- [143] 도 12는 본 개시의 일 실시 예에 따른 전자 장치의 제어 방법을 설명하기 위한 흐름도이다.
- [144] 먼저, 제1 학습 데이터를 통해 학습된 제1 인공 지능 모델 및 제1 학습 데이터를 통해 학습된 제2 인공 지능 모델 각각에 제2 학습 데이터를 입력한다(S1210). 그리고, 제1 인공 지능 모델에 포함된 복수의 제1 레이어 각각의 출력 및 제2 인공 지능 모델에 포함된 복수의 제2 레이어 각각의 출력에 기초하여 제2 인공 지능 모델을 재학습한다(S1220). 여기서, 복수의 제1 레이어 각각은 복수의 2차원 필터를 포함하고, 복수의 제2 레이어 각각은 대응되는 제1 레이어의 복수의 2차원 필터 각각의 크기가 축소된 복수의 필터를 포함할 수 있다.
- [145] 또한, 복수의 제1 레이어의 개수는 복수의 제2 레이어의 개수와 동일하고, 재학습하는 단계(S1220)는 복수의 제1 레이어 각각의 출력을 대응되는 제2 레이어의 출력과 비교하여 복수의 비교 결과를 획득하고, 복수의 비교 결과에 기초하여 제2 인공 지능 모델을 재학습할 수 있다.
- [146] 여기서, 복수의 제1 레이어 각각에 포함된 복수의 2차원 필터의 개수는 대응되는 제2 레이어에 포함된 복수의 필터의 개수와 동일하며, 복수의 제1 레이어 각각의 출력의 크기는 대응되는 제2 레이어의 출력의 크기와 동일할 수 있다.
- [147] 그리고, 재학습하는 단계(S1220)는 복수의 제2 레이어 각각의 출력을 제3 인공 지능 모델 각각에 입력하고, 제3 인공 지능 모델로부터 출력되는 복수의 제2 레이어의 출력에 대한 복수의 구별 결과를 획득하며, 복수의 비교 결과 및 복수의 구별 결과에 기초하여 제2 인공 지능 모델을 재학습하고, 제3 인공 지능 모델은 복수의 제1 레이어 각각의 출력 및 복수의 제2 레이어 각각의 출력을 구별하도록 학습된 인공 지능 모델일 수 있다.
- [148] 여기서, 제2 학습 데이터를 재학습된 제2 인공 지능 모델 각각에 입력하는 단계 및 복수의 제1 레이어 각각의 출력 및 재학습된 제2 인공 지능 모델에 포함된 복수의 제3 레이어 각각의 출력을 구별하도록 제3 인공 지능 모델을 재학습하는 단계를 더 포함할 수 있다.
- [149] 또한, 재학습하는 단계(S1220)는 복수의 비교 결과 및 복수의 구별 결과를 가중합하여 제2 인공 지능 모델을 재학습할 수 있다.
- [150] 여기서, 재학습하는 단계(S1220)는 복수의 제1 레이어 중 마지막 레이어에 대응되는 비교 결과의 가중치를 나머지 레이어에 대응되는 복수의 비교 결과의 가중치 이상으로 부여하여 제2 인공 지능 모델을 재학습할 수 있다.
- [151] 한편, 복수의 필터 각각은 $1 \times N$ 형태 또는 $N \times 1$ 형태이고, 입력 데이터를 행(row) 단위 또는 컬럼(column) 단위로 독출하는 단계 및 독출된 입력 데이터를 재학습된 제2 인공 지능 모델에 입력하여 입력 데이터를 처리하는 단계를 더 포함할 수 있다.
- [152] 그리고, 제1 인공 지능 모델 및 제2 인공 지능 모델은 컨볼루션 신경망(Convolutional Neural Network, CNN) 형태로 구현될 수 있다.

- [153] 이상과 같은 본 개시의 다양한 실시 예에 따르면, 전자 장치는 Teacher network 및 Student network의 각 레이어의 출력을 비교하여 Student network를 재학습함에 따라 로우-레벨(low-level)의 특징이 부각된 인공 지능 모델을 획득할 수 있다.
- [154] 또한, 전자 장치는 Teacher network 및 Student network의 각 레이어의 출력을 구분하도록 학습된 discriminator를 더 이용하여 로우-레벨의 특징이 좀더 부각된 인공 지능 모델을 획득할 수도 있다.
- [155] 그리고, 전자 장치는 Teacher network의 2차원 형태의 필터의 크기를 축소한 Student network를 이용함에 따라 인공 지능 모델의 용량을 경량화하고 on the fly 방식으로 입력 데이터를 처리할 수 있다.
- [156] 한편, 본 개시의 일 실시 예에 따르면, 이상에서 설명된 다양한 실시 예들은 기기(machine)(예: 컴퓨터)로 읽을 수 있는 저장 매체(machine-readable storage media)에 저장된 명령어를 포함하는 소프트웨어로 구현될 수 있다. 기기는, 저장 매체로부터 저장된 명령어를 호출하고, 호출된 명령어에 따라 동작이 가능한 장치로서, 개시된 실시 예들에 따른 전자 장치(예: 전자 장치(A))를 포함할 수 있다. 명령이 프로세서에 의해 실행될 경우, 프로세서가 직접, 또는 프로세서의 제어 하에 다른 구성요소들을 이용하여 명령에 해당하는 기능을 수행할 수 있다. 명령은 컴파일러 또는 인터프리터에 의해 생성 또는 실행되는 코드를 포함할 수 있다. 기기로 읽을 수 있는 저장매체는, 비일시적(non-transitory) 저장매체의 형태로 제공될 수 있다. 여기서, '비일시적'은 저장매체가 신호(signal)를 포함하지 않으며 실재(tangible)하다는 것을 의미할 뿐 데이터가 저장매체에 반영구적 또는 임시적으로 저장됨을 구분하지 않는다.
- [157] 또한, 본 개시의 일 실시 예에 따르면, 이상에서 설명된 다양한 실시 예들에 따른 방법은 컴퓨터 프로그램 제품(computer program product)에 포함되어 제공될 수 있다. 컴퓨터 프로그램 제품은 상품으로서 판매자 및 구매자 간에 거래될 수 있다. 컴퓨터 프로그램 제품은 기기로 읽을 수 있는 저장 매체(예: compact disc read only memory (CD-ROM))의 형태로, 또는 어플리케이션 스토어(예: 플레이 스토어™)를 통해 온라인으로 배포될 수 있다. 온라인 배포의 경우에, 컴퓨터 프로그램 제품의 적어도 일부는 제조사의 서버, 어플리케이션 스토어의 서버, 또는 중계 서버의 메모리와 같은 저장 매체에 적어도 일시 저장되거나, 임시적으로 생성될 수 있다.
- [158] 또한, 본 개시의 일 실시 예에 따르면, 이상에서 설명된 다양한 실시 예들은 소프트웨어(software), 하드웨어(hardware) 또는 이들의 조합을 이용하여 컴퓨터(computer) 또는 이와 유사한 장치로 읽을 수 있는 기록 매체 내에서 구현될 수 있다. 일부 경우에 있어 본 명세서에서 설명되는 실시 예들이 프로세서 자체로 구현될 수 있다. 소프트웨어적인 구현에 의하면, 본 명세서에서 설명되는 절차 및 기능과 같은 실시 예들은 별도의 소프트웨어 모듈들로 구현될 수 있다. 소프트웨어 모듈들 각각은 본 명세서에서 설명되는 하나 이상의 기능 및 동작을 수행할 수 있다.

- [159] 한편, 상술한 다양한 실시 예들에 따른 기기의 프로세싱 동작을 수행하기 위한 컴퓨터 명령어(computer instructions)는 비일시적 컴퓨터 판독 가능 매체(non-transitory computer-readable medium)에 저장될 수 있다. 이러한 비일시적 컴퓨터 판독 가능 매체에 저장된 컴퓨터 명령어는 특정 기기의 프로세서에 의해 실행되었을 때 상술한 다양한 실시 예에 따른 기기에서의 처리 동작을 특정 기기가 수행하도록 한다. 비일시적 컴퓨터 판독 가능 매체란 레지스터, 캐쉬, 메모리 등과 같이 짧은 순간 동안 데이터를 저장하는 매체가 아니라 반영구적으로 데이터를 저장하며, 기기에 의해 판독(reading)이 가능한 매체를 의미한다. 비일시적 컴퓨터 판독 가능 매체의 구체적인 예로는, CD, DVD, 하드 디스크, 블루레이 디스크, USB, 메모리카드, ROM 등이 있을 수 있다.
- [160] 또한, 상술한 다양한 실시 예들에 따른 구성 요소(예: 모듈 또는 프로그램) 각각은 단위 또는 복수의 개체로 구성될 수 있으며, 전술한 해당 서브 구성 요소들 중 일부 서브 구성 요소가 생략되거나, 또는 다른 서브 구성 요소가 다양한 실시 예에 더 포함될 수 있다. 대체적으로 또는 추가적으로, 일부 구성 요소들(예: 모듈 또는 프로그램)은 하나의 개체로 통합되어, 통합되기 이전의 각각의 해당 구성 요소에 의해 수행되는 기능을 동일 또는 유사하게 수행할 수 있다. 다양한 실시예들에 따른, 모듈, 프로그램 또는 다른 구성 요소에 의해 수행되는 동작들은 순차적, 병렬적, 반복적 또는 휴리스틱하게 실행되거나, 적어도 일부 동작이 다른 순서로 실행되거나, 생략되거나, 또는 다른 동작이 추가될 수 있다.
- [161] 이상에서는 본 개시의 바람직한 실시 예에 대하여 도시하고 설명하였지만, 본 개시는 상술한 특정의 실시 예에 한정되지 아니하며, 청구범위에서 청구하는 본 개시의 요지를 벗어남이 없이 당해 개시에 속하는 기술분야에서 통상의 지식을 가진 자에 의해 다양한 변형실시가 가능한 것은 물론이고, 이러한 변형실시들은 본 개시의 기술적 사상이나 전망으로부터 개별적으로 이해되어져서는 안될 것이다.

청구범위

- [청구항 1] 전자 장치에 있어서,
제1 학습 데이터를 통해 학습된 제1 인공 지능 모델에 대한 정보 및 상기 제1 학습 데이터를 통해 학습된 제2 인공 지능 모델에 대한 정보가 저장된 메모리; 및
상기 메모리와 연결되어 상기 전자 장치를 제어하는 프로세서;를 포함하며,
상기 프로세서는,
제2 학습 데이터를 상기 제1 인공 지능 모델 및 상기 제2 인공 지능 모델 각각에 입력하고,
상기 제1 인공 지능 모델에 포함된 복수의 제1 레이어 각각의 출력 및 상기 제2 인공 지능 모델에 포함된 복수의 제2 레이어 각각의 출력에 기초하여 상기 제2 인공 지능 모델을 재학습하며,
상기 복수의 제1 레이어 각각은, 복수의 2차원 필터를 포함하고,
상기 복수의 제2 레이어 각각은, 대응되는 제1 레이어의 복수의 2차원 필터 각각의 크기가 축소된 복수의 필터를 포함하는, 전자 장치.
- [청구항 2] 제1항에 있어서,
상기 복수의 제1 레이어의 개수는, 상기 복수의 제2 레이어의 개수와 동일하고,
상기 프로세서는,
상기 복수의 제1 레이어 각각의 출력을 대응되는 제2 레이어의 출력과 비교하여 복수의 비교 결과를 획득하고,
상기 복수의 비교 결과에 기초하여 상기 제2 인공 지능 모델을 재학습하는, 전자 장치.
- [청구항 3] 제2항에 있어서,
상기 복수의 제1 레이어 각각에 포함된 상기 복수의 2차원 필터의 개수는, 대응되는 제2 레이어에 포함된 상기 복수의 필터의 개수와 동일하며,
상기 복수의 제1 레이어 각각의 출력의 크기는, 대응되는 제2 레이어의 출력의 크기와 동일한, 전자 장치.
- [청구항 4] 제3항에 있어서,
상기 메모리는,
상기 복수의 제1 레이어 각각의 출력 및 상기 복수의 제2 레이어 각각의 출력을 구별하도록 학습된 제3 인공 지능 모델에 대한 정보를 저장하며,
상기 프로세서는,
상기 복수의 제2 레이어 각각의 출력을 상기 제3 인공 지능 모델 각각에 입력하고,
상기 제3 인공 지능 모델로부터 출력되는 상기 복수의 제2 레이어의

출력에 대한 복수의 구별 결과를 획득하며,
 상기 복수의 비교 결과 및 상기 복수의 구별 결과에 기초하여 상기 제2
 인공 지능 모델을 재학습하는, 전자 장치.

[청구항 5] 제4항에 있어서,
 상기 프로세서는,
 상기 제2 학습 데이터를 상기 재학습된 제2 인공 지능 모델 각각에
 입력하고,
 상기 복수의 제1 레이어 각각의 출력 및 상기 재학습된 제2 인공 지능
 모델에 포함된 복수의 제3 레이어 각각의 출력을 구별하도록 상기 제3
 인공 지능 모델을 재학습하는, 전자 장치.

[청구항 6] 제4항에 있어서,
 상기 프로세서는,
 상기 복수의 비교 결과 및 상기 복수의 구별 결과를 가중합하여 상기 제2
 인공 지능 모델을 재학습하는, 전자 장치.

[청구항 7] 제6항에 있어서,
 상기 프로세서는,
 상기 복수의 제1 레이어 중 마지막 레이어에 대응되는 비교 결과의
 가중치를 나머지 레이어에 대응되는 복수의 비교 결과의 가중치
 이상으로 부여하여 상기 제2 인공 지능 모델을 재학습하는, 전자 장치.

[청구항 8] 제1항에 있어서,
 상기 복수의 필터 각각은,
 $1 \times N$ 형태 또는 $N \times 1$ 형태이고,
 상기 프로세서는,
 상기 메모리로부터 입력 데이터를 행(row) 단위 또는 컬럼(column) 단위로
 독출하고,
 상기 독출된 입력 데이터를 상기 재학습된 제2 인공 지능 모델에
 입력하여 상기 입력 데이터를 처리하는, 전자 장치.

[청구항 9] 제1항에 있어서,
 상기 제1 인공 지능 모델 및 상기 제2 인공 지능 모델은,
 컨볼루션 신경망(Convolutional Neural Network, CNN) 형태로 구현된,
 전자 장치.

[청구항 10] 제1항에 있어서,
 상기 메모리는, 상기 제1 학습 데이터를 통해 학습된 보조 인공 지능
 모델에 대한 정보를 더 저장하고,
 상기 프로세서는,
 상기 제2 학습 데이터를 상기 보조 인공 지능 모델에 입력하고,
 상기 복수의 제1 레이어 각각의 출력 및 상기 보조 인공 지능 모델에
 포함된 복수의 보조 레이어 각각의 출력에 기초하여 상기 보조 인공 지능

모델을 재학습하며,
 상기 제2 학습 데이터를 상기 재학습된 보조 인공 지능 모델에 입력하고,
 상기 재학습된 보조 인공 지능 모델에 포함된 복수의 보조 레이어 각각의 출력 및 상기 복수의 제2 레이어 각각의 출력에 기초하여 상기 제2 인공 지능 모델을 재학습하며,
 상기 복수의 보조 레이어 각각은, 대응되는 제1 레이어의 복수의 2차원 필터 각각의 크기가 축소된 복수의 필터를 포함하는, 전자 장치.

[청구항 11]

제10항에 있어서,
 상기 메모리는, 상기 제1 학습 데이터를 통해 학습된 제1 보조 인공 지능 모델에 대한 정보 및 상기 제1 학습 데이터를 통해 학습된 제2 보조 인공 지능 모델에 대한 정보를 더 저장하고,
 상기 프로세서는,
 상기 제2 학습 데이터를 상기 제1 보조 인공 지능 모델 및 상기 제2 보조 인공 지능 모델에 입력하고,
 상기 복수의 제1 레이어 각각의 출력 및 상기 제1 보조 인공 지능 모델에 포함된 복수의 제1 보조 레이어 각각의 출력에 기초하여 상기 제1 보조 인공 지능 모델을 재학습하고, 상기 복수의 제1 레이어 각각의 출력 및 상기 제2 보조 인공 지능 모델에 포함된 복수의 제2 보조 레이어 각각의 출력에 기초하여 상기 제2 보조 인공 지능 모델을 재학습하며,
 상기 제2 학습 데이터를 상기 재학습된 제1 보조 인공 지능 모델 및 상기 재학습된 제2 보조 인공 지능 모델에 입력하고,
 상기 재학습된 제1 보조 인공 지능 모델에 포함된 복수의 제1 보조 레이어 각각의 출력, 상기 재학습된 제2 보조 인공 지능 모델에 포함된 복수의 제2 보조 레이어 각각의 출력 및 상기 복수의 제2 레이어 각각의 출력에 기초하여 상기 제2 인공 지능 모델을 재학습하며,
 상기 복수의 제1 보조 레이어 각각은, 대응되는 제1 레이어의 복수의 2차원 필터 각각의 크기가 축소된 복수의 필터를 포함하고,
 상기 복수의 제2 보조 레이어 각각은, 대응되는 제1 레이어의 복수의 2차원 필터 각각의 크기가 축소된 복수의 필터를 포함하고,
 상기 복수의 제1 보조 레이어 각각에 포함된 복수의 필터 각각은, 대응되는 제2 보조 레이어 각각에 포함된 필터와 형태가 상이한, 전자 장치.

[청구항 12]

전자 장치의 제어 방법에 있어서,
 제1 학습 데이터를 통해 학습된 제1 인공 지능 모델 및 상기 제1 학습 데이터를 통해 학습된 제2 인공 지능 모델 각각에 제2 학습 데이터를 입력하는 단계; 및
 상기 제1 인공 지능 모델에 포함된 복수의 제1 레이어 각각의 출력 및 상기 제2 인공 지능 모델에 포함된 복수의 제2 레이어 각각의 출력에

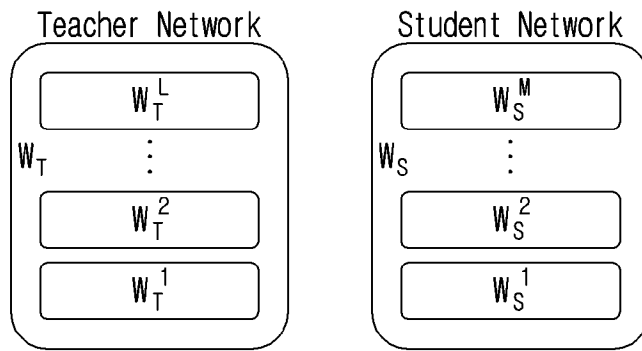
기초하여 상기 제2 인공 지능 모델을 재학습하는 단계;를 포함하며,
 상기 복수의 제1 레이어 각각은, 복수의 2차원 필터를 포함하고,
 상기 복수의 제2 레이어 각각은, 대응되는 제1 레이어의 복수의 2차원
 필터 각각의 크기가 축소된 복수의 필터를 포함하는, 제어 방법.

[청구항 13] 제12항에 있어서,
 상기 복수의 제1 레이어의 개수는, 상기 복수의 제2 레이어의 개수와
 동일하고,
 상기 재학습하는 단계는,
 상기 복수의 제1 레이어 각각의 출력을 대응되는 제2 레이어의 출력과
 비교하여 복수의 비교 결과를 획득하고,
 상기 복수의 비교 결과에 기초하여 상기 제2 인공 지능 모델을
 재학습하는, 제어 방법.

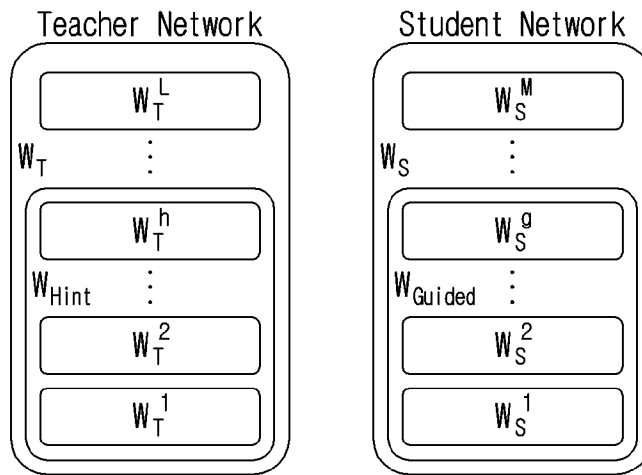
[청구항 14] 제13항에 있어서,
 상기 복수의 제1 레이어 각각에 포함된 상기 복수의 2차원 필터의 개수는,
 대응되는 제2 레이어에 포함된 상기 복수의 필터의 개수와 동일하며,
 상기 복수의 제1 레이어 각각의 출력의 크기는, 대응되는 제2 레이어의
 출력의 크기와 동일한, 제어 방법.

[청구항 15] 제14항에 있어서,
 상기 재학습하는 단계는,
 상기 복수의 제2 레이어 각각의 출력을 제3 인공 지능 모델 각각에
 입력하고,
 상기 제3 인공 지능 모델로부터 출력되는 상기 복수의 제2 레이어의
 출력에 대한 복수의 구별 결과를 획득하며,
 상기 복수의 비교 결과 및 상기 복수의 구별 결과에 기초하여 상기 제2
 인공 지능 모델을 재학습하고,
 상기 제3 인공 지능 모델은,
 상기 복수의 제1 레이어 각각의 출력 및 상기 복수의 제2 레이어 각각의
 출력을 구별하도록 학습된 인공 지능 모델인, 제어 방법.

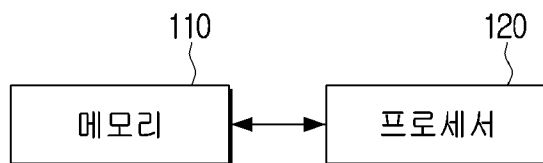
[도 1a]



[도 1b]



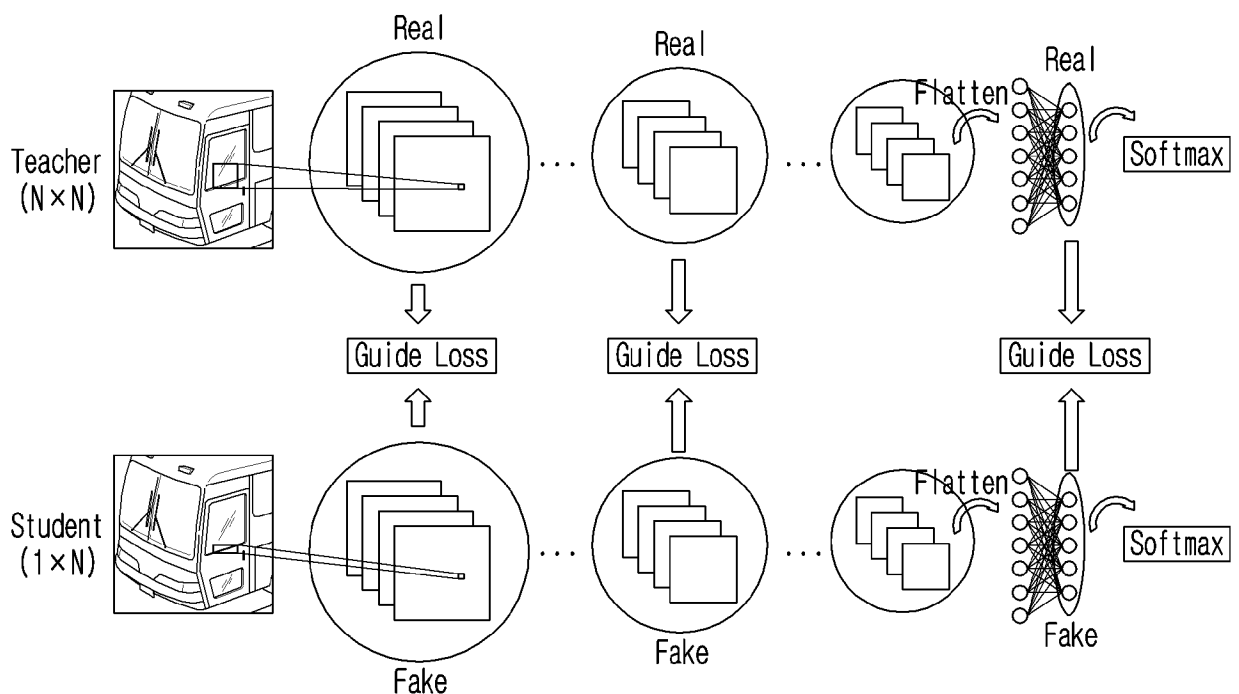
[도 2]

100

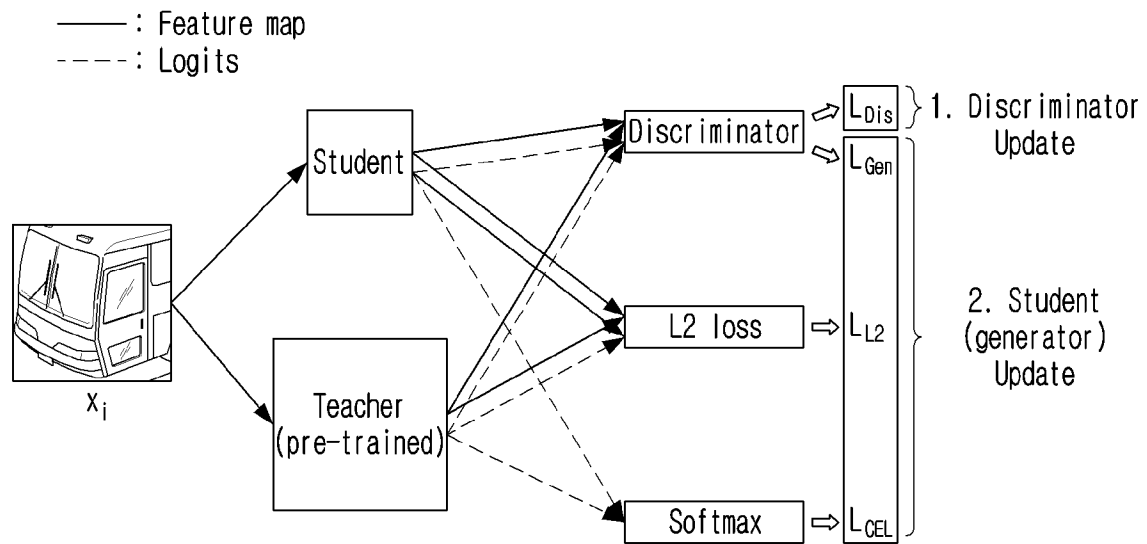
[도3]

Teacher Configuration	Student Configuration
Input (32 × 32 RGB image)	Input (32 × 32 RGB image)
Conv 3 × 3 - 64	Conv 1 × 3 - 64
Guide Loss	
BatchNorm	
Conv 3 × 3 - 64	Conv 1 × 3 - 64
BatchNorm, MaxPool	
Conv 3 × 3 - 128	Conv 1 × 3 - 128
Guide Loss	
BatchNorm	
Conv 3 × 3 - 128	Conv 1 × 3 - 128
BatchNorm, MaxPool	
Conv 3 × 3 - 256	Conv 1 × 3 - 256
Guide Loss	
BatchNorm	
Conv 3 × 3 - 256	Conv 1 × 3 - 256
BatchNorm, MaxPool	
Conv 3 × 3 - 512	Conv 1 × 3 - 512
Guide Loss	
BatchNorm	
Conv 3 × 3 - 512	Conv 1 × 3 - 512
BatchNorm, MaxPool	
FC layer 10	
Teacher outputs	Student outputs
Guide Loss	
Targets	

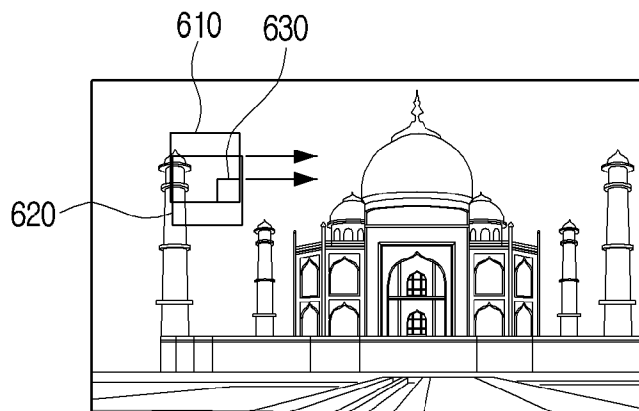
[도4]



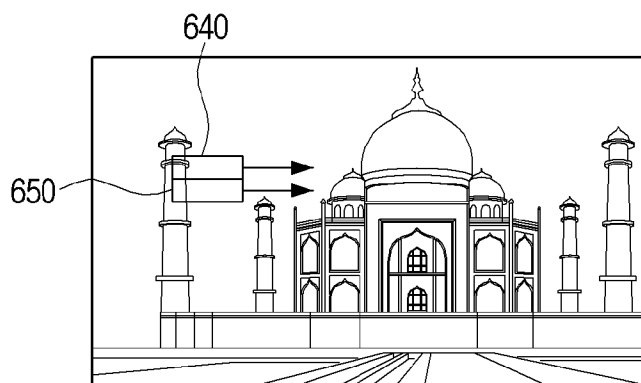
[도5]



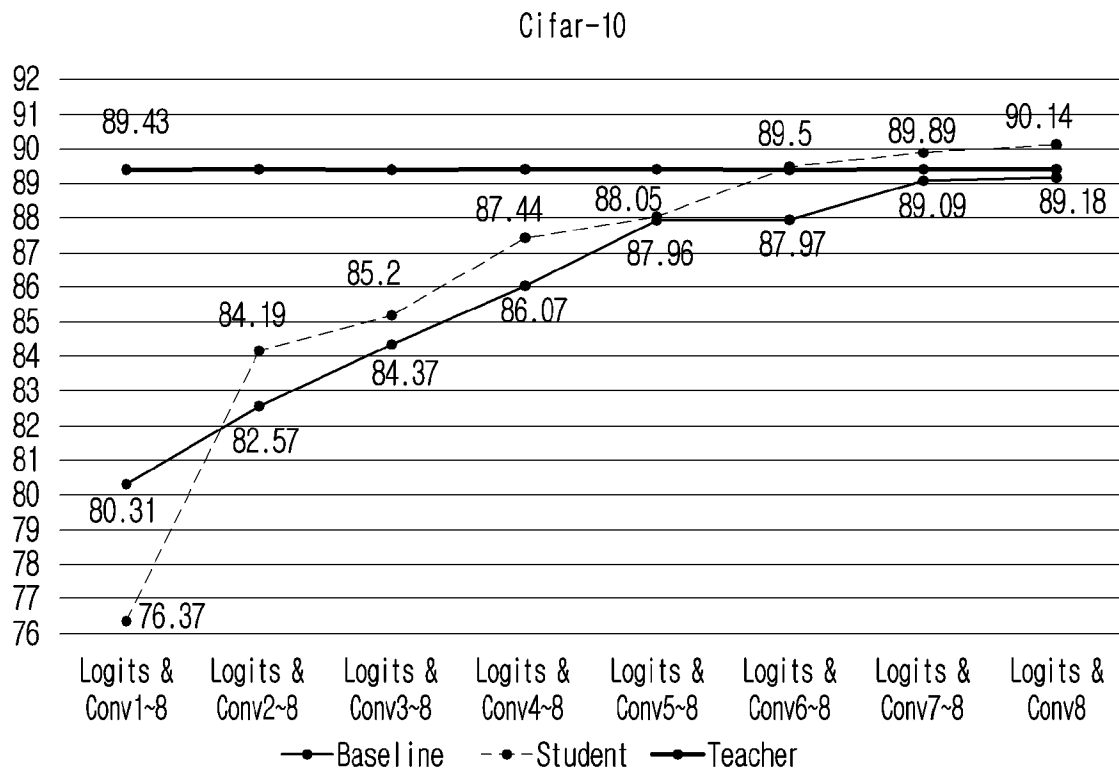
[도6a]



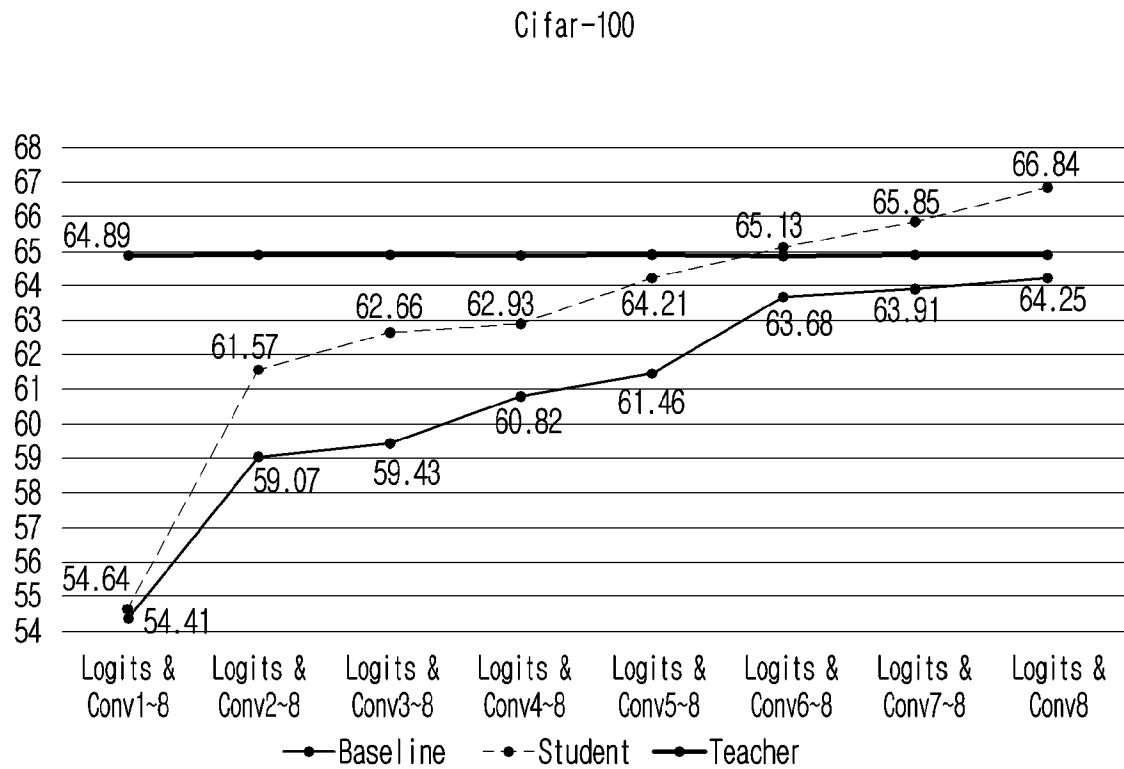
[도6b]



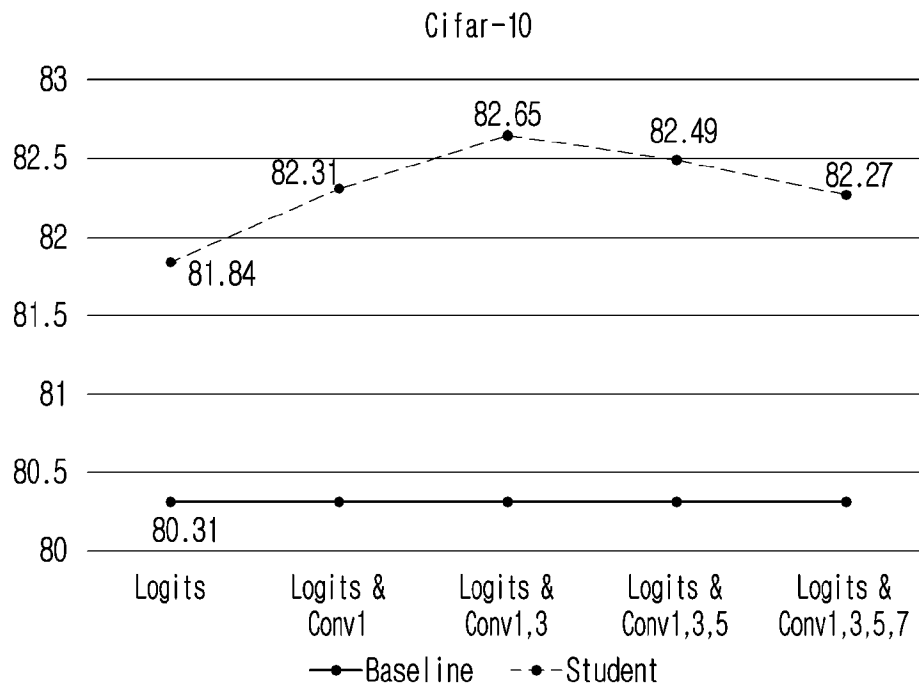
[도7a]



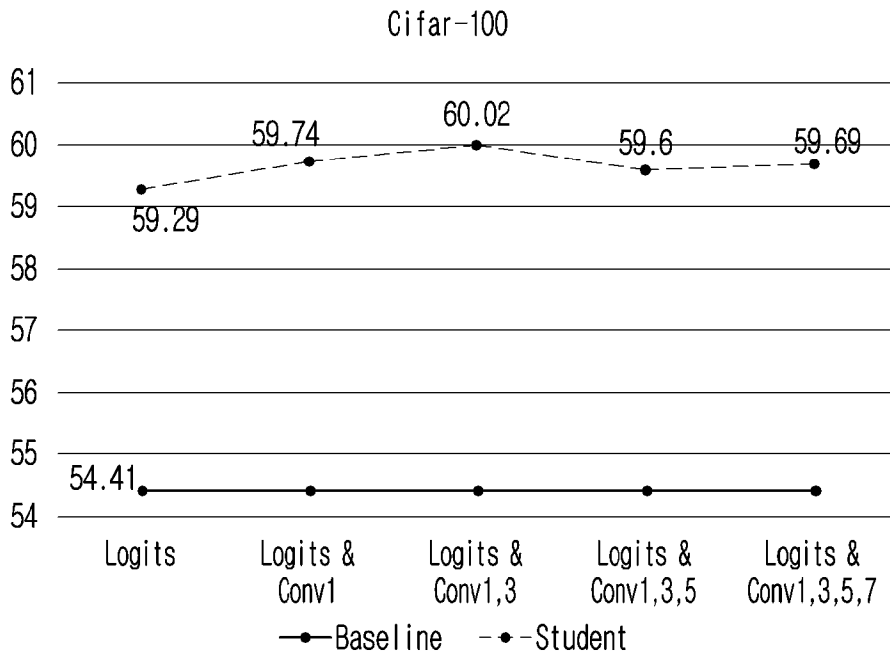
[도7b]



[도8a]



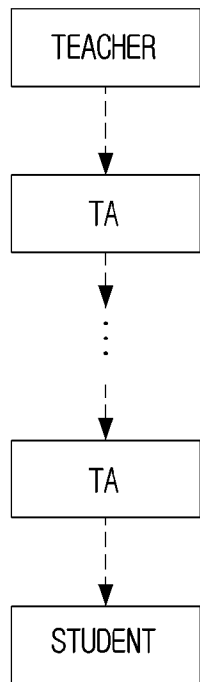
[도8b]



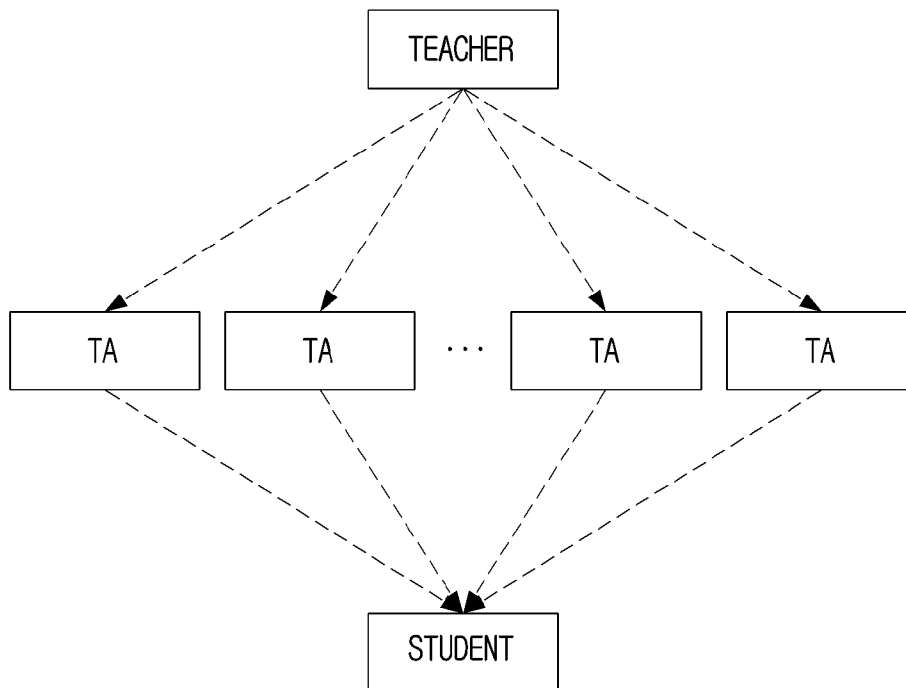
[도9]

Loss composition	CIFAR - 10	CIFAR - 100
Baseline	80.31	54.41
$L_{HT} (T=10)$	84.39	61.95
$L_{HT} + L_{L2}$	84.54	62.26
$L_{HT} + L_{Gen}$	84.6	62.41
$L_{HT} + L_{L2} + L_{Gen}$ (ours)	84.76	62.55

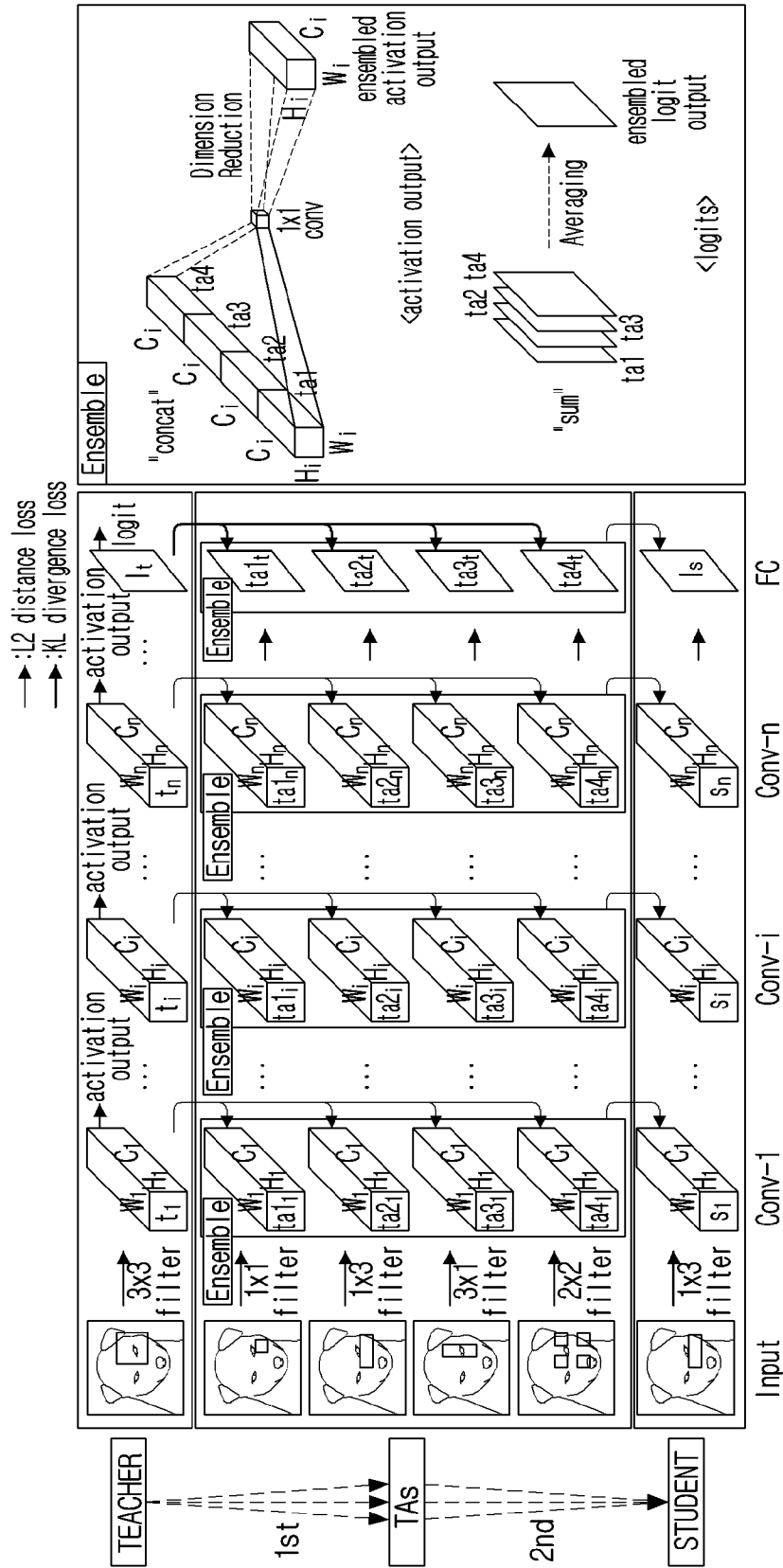
[도10]



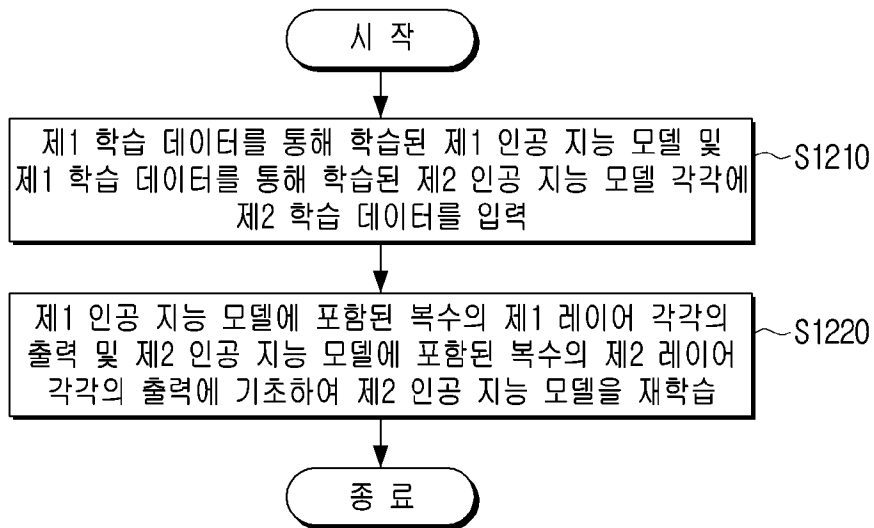
[도11a]



[도11b]



[도12]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/KR2021/000891

A. CLASSIFICATION OF SUBJECT MATTER		
G06N 3/04(2006.01)i; G06N 3/08(2006.01)i; G06N 20/00(2019.01)i; G06N 5/02(2006.01)i		
According to International Patent Classification (IPC) or to both national classification and IPC		
B. FIELDS SEARCHED		
Minimum documentation searched (classification system followed by classification symbols) G06N 3/04(2006.01); G06N 3/08(2006.01); G06N 99/00(2010.01); G06T 3/40(2006.01)		
Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched Korean utility models and applications for utility models: IPC as above Japanese utility models and applications for utility models: IPC as above		
Electronic data base consulted during the international search (name of data base and, where practicable, search terms used) eKOMPASS (KIPO internal) & keywords: 지식 추출(knowledge distillation), 교사(teacher), 학생(student), 레이어(layer), 학습(learning), 필터(filter), 가중합(weighted sum)		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	YIM, Junho et al. A Gift from Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 713 0-7138, 26 July 2017. See sections 3-4; and figure 2.	1-15
Y	JP 2019-067078 A (UNIV. TSUKUBA) 25 April 2019 (2019-04-25) See paragraph [0026]; and claim 1.	1-15
Y	KR 10-2019-0068255 A (SAMSUNG ELECTRONICS CO., LTD.) 18 June 2019 (2019-06-18) See claim 6.	6-7
A	US 2018-0365564 A1 (TUSIMPLE) 20 December 2018 (2018-12-20) See paragraphs [0037]-[0064]; and figures 1-2.	1-15
A	KR 10-2017-0034258 A (SAMSUNG ELECTRONICS CO., LTD.) 28 March 2017 (2017-03-28) See paragraphs [0090]-[0113]; and figures 6-7.	1-15
☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.		
* Special categories of cited documents: "A" document defining the general state of the art which is not considered to be of particular relevance "D" document cited by the applicant in the international application "E" earlier application or patent but published on or after the international filing date "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) "O" document referring to an oral disclosure, use, exhibition or other means "P" document published prior to the international filing date but later than the priority date claimed "T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention "X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone "Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art "&" document member of the same patent family		
Date of the actual completion of the international search 04 May 2021		Date of mailing of the international search report 06 May 2021
Name and mailing address of the ISA/KR Korean Intellectual Property Office Government Complex-Daejeon Building 4, 189 Cheongsaro, Seo-gu, Daejeon 35208 Facsimile No. +82-42-481-8578		Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/KR2021/000891

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
JP	2019-067078	A	25 April 2019	None			
KR	10-2019-0068255	A	18 June 2019	CN	110020717	A	16 July 2019
				US	2019-0180177	A1	13 June 2019
US	2018-0365564	A1	20 December 2018	CN	107247989	A	13 October 2017
				CN	107247989	B	24 November 2020
KR	10-2017-0034258	A	28 March 2017	CN	106548190	A	29 March 2017
				EP	3144859	A2	22 March 2017
				EP	3144859	A3	14 June 2017
				US	10410114	B2	10 September 2019
				US	2017-0083829	A1	23 March 2017

A. 발명이 속하는 기술분류(국제특허분류(IPC)) G06N 3/04(2006.01)i; G06N 3/08(2006.01)i; G06N 20/00(2019.01)i; G06N 5/02(2006.01)i		
B. 조사된 분야 조사된 최소문헌(국제특허분류를 기재) G06N 3/04(2006.01); G06N 3/08(2006.01); G06N 99/00(2010.01); G06T 3/40(2006.01) 조사된 기술분야에 속하는 최소문헌 이외의 문헌 한국등록실용신안공보 및 한국공개실용신안공보: 조사된 최소문헌란에 기재된 IPC 일본등록실용신안공보 및 일본공개실용신안공보: 조사된 최소문헌란에 기재된 IPC 국제조사에 이용된 전산 데이터베이스(데이터베이스의 명칭 및 검색어(해당하는 경우)) eKOMPASS(특허청 내부 검색시스템) & 키워드: 지식 추출(knowledge distillation), 교사(teacher), 학생(student), 레이어(layer), 학습(learning), 필터(filter), 가중합(weighted sum)		
C. 관련 문헌		
카테고리*	인용문헌명 및 관련 구절(해당하는 경우)의 기재	관련 청구항
Y	JUNHO YIM 등, 'A Gift from Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning', 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 페이지 7130-7138, 2017.07.26 섹션 3-4; 및 도면 2	1-15
Y	JP 2019-067078 A (UNIV. TSUKUBA) 2019.04.25 단락 [0026]; 및 청구항 1	1-15
Y	KR 10-2019-0068255 A (삼성전자주식회사) 2019.06.18 청구항 6	6-7
A	US 2018-0365564 A1 (TUSIMPLE) 2018.12.20 단락 [0037]-[0064]; 및 도면 1-2	1-15
A	KR 10-2017-0034258 A (삼성전자주식회사) 2017.03.28 단락 [0090]-[0113]; 및 도면 6-7	1-15
☐ 추가 문헌이 C(계속)에 기재되어 있습니다. ☒ 대응특허에 관한 별지를 참조하십시오.		
* 인용된 문헌의 특별 카테고리: "A" 특별히 관련이 없는 것으로 보이는 일반적인 기술수준을 정의한 문헌 "D" 본 국제출원에서 출원인이 인용한 문헌 "E" 국제출원일보다 빠른 출원일 또는 우선일을 가지나 국제출원일 이후에 공개된 선출원 또는 특허 문헌 "L" 우선권 주장에 의문을 제기하는 문헌 또는 다른 인용문헌의 공개일 또는 다른 특별한 이유(이유를 명시)를 밝히기 위하여 인용된 문헌 "O" 구두 개시, 사용, 전시 또는 기타 수단을 언급하고 있는 문헌 "P" 우선일 이후에 공개되었으나 국제출원일 이전에 공개된 문헌 "T" 국제출원일 또는 우선일 후에 공개된 문헌으로, 출원과 상충하지 않으며 발명의 기초가 되는 원리나 이론을 이해하기 위해 인용된 문헌 "X" 특별한 관련이 있는 문헌. 해당 문헌 하나만으로 청구된 발명의 신규성 또는 진보성이 없는 것으로 본다. "Y" 특별한 관련이 있는 문헌. 해당 문헌이 하나 이상의 다른 문헌과 조합하는 경우로 그 조합이 당업자에게 자명한 경우 청구된 발명은 진보성이 없는 것으로 본다. "&" 동일한 대응특허문헌에 속하는 문헌		
국제조사의 실제 완료일 2021년05월04일(04.05.2021)		국제조사보고서 발송일 2021년05월06일(06.05.2021)
ISA/KR의 명칭 및 우편주소 대한민국 특허청 (35208) 대전광역시 서구 청사로 189, 4동 (둔산동, 정부대전청사) 팩스 번호 +82-42-481-8578		심사관 양정록 전화번호 +82-42-481-5709

국 제 조 사 보 고 서
대응특허에 관한 정보

국제출원번호

PCT/KR2021/000891

국제조사보고서에서 인용된 특허문헌	공개일	대응특허문헌	공개일
JP 2019-067078 A	2019/04/25	없음	
KR 10-2019-0068255 A	2019/06/18	CN 110020717 A	2019/07/16
		US 2019-0180177 A1	2019/06/13
US 2018-0365564 A1	2018/12/20	CN 107247989 A	2017/10/13
		CN 107247989 B	2020/11/24
KR 10-2017-0034258 A	2017/03/28	CN 106548190 A	2017/03/29
		EP 3144859 A2	2017/03/22
		EP 3144859 A3	2017/06/14
		US 10410114 B2	2019/09/10
		US 2017-0083829 A1	2017/03/23