(12) 특허협력조약에 의하여 공개된 국제출원

(19) 세계지식재산권기구
국제사무국

(43) 국제공개일
2023 년 6 월 22 일 (22.06.2023)

WIPO | PCT

(10) 국제공개번호
WO 2023/113372 A1

(51) 국제특허분류:
G06N 3/08 (2006.01) *G06N 3/04* (2006.01)

(21) 국제출원번호: PCT/KR2022/019879

(22) 국제출원일: 2022 년 12 월 8 일 (08.12.2022)

(25) 출원언어: 한국어

(26) 공개언어: 한국어

(30) 우선권정보:
10-2021-0180528 2021 년 12 월 16 일 (16.12.2021) KR

(71) 출원인: 창원대학교 산학협력단 (**CHANGWON NATIONAL UNIVERSITY INDUSTRY ACADEMY COOPERATION CORPS**) [KR/KR]; 51140 경상남도 창원시 의창구 창원대학로 20 (퇴촌동), Gyeongsangnam-do (KR).

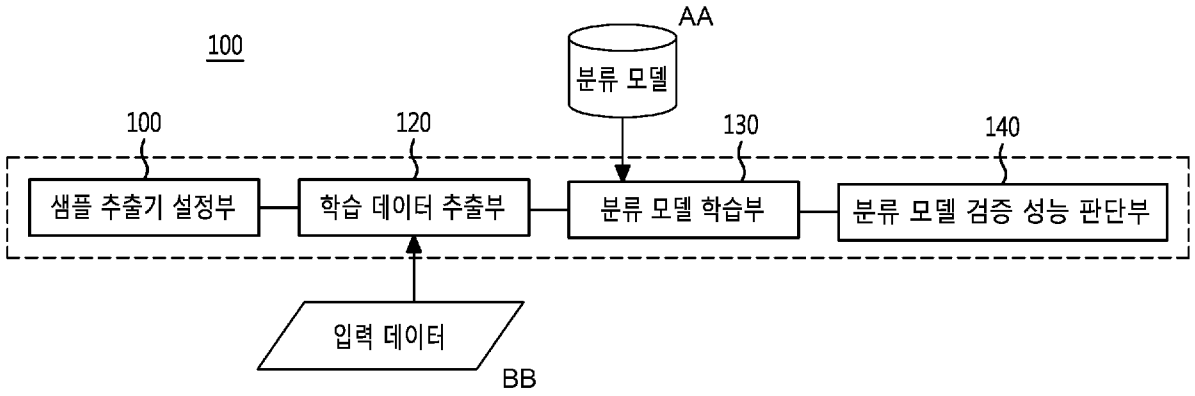
(72) 발명자: 차정원 (**CHA, Jeong-Won**); 51498 경상남도 창원시 성산구 동산로 115, 105동 1407호 (상남동), Gyeongsangnam-do (KR). 성수진 (**SEONG, Su-Jin**); 51244 경상남도 창원시 마산회원구 내서읍 광려천동로 22, 512호, Gyeongsangnam-do (KR).

(74) 대리인: 김종선 (**KIM, Jong Sun**); 06242 서울특별시 강남구 역삼로3길 11, 12층 (역삼동, 광성빌딩), 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, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, 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,

(54) **Title:** APPARATUS AND METHOD FOR LABEL-BASED SAMPLE EXTRACTION FOR IMPROVEMENT OF DEEP LEARNING CLASSIFICATION MODEL PERFORMANCE FOR IMBALANCED DATA

(54) **발명의 명칭:** 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 장치 및 그 방법



100 ... Sample extractor configuration unit
120 ... Training data extraction unit
130 ... Classification model training unit
140 ... Classification model verification performance determining unit
AA ... Input data
BB ... Classification model

(57) **Abstract:** The present invention relates to an apparatus and a method for label-based sample extraction for improvement of deep learning classification model performance for imbalanced data, the apparatus comprising: a sample extractor configuration unit for configuring one of a first sample extractor on the basis of the number of label samples of entire training data and a second sample extractor on the basis of the number of correct label samples existing in the entire training data; a training data extraction unit for configuring sample training data by extracting a sample to be used for training a classification model from the entire training data by using the configured first sample extractor or second sample extractor; and a classification model training unit for training and updating the classification model by using the configured sample training data.

WO 2023/113372 A1

[다음 쪽 계속]

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 역내 권리의 보호를 위하여): ARIPO (BW, CV, 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, ME, 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).

규칙 4.17에 의한 선언서:

— 신규성을 해치지 아니하는 개시 또는 신규성 상실의 예외에 관한 선언 (규칙 4.17(v))

공개:

— 국제조사보고서와 함께 (조약 제21조(3))

(57) 요약서: 본 발명은 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 장치 및 그 방법에 관한 것으로, 전체 학습 데이터의 레이블 샘플수를 기반으로 하는 제1 샘플 추출기 및 전체 학습 데이터에 나타나는 정답 레이블 샘플수를 기반으로 하는 제2 샘플 추출기 중 하나를 설정하는 샘플 추출기 설정부, 설정된 제1 샘플 추출기 또는 제2 샘플 추출기를 이용하여 전체 학습 데이터에서 분류 모델의 학습에 사용할 샘플을 추출하여 샘플 학습 데이터를 구성하는 학습 데이터 추출부 및 구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습하고 업데이트하는 분류 모델 학습부를 포함한다.

명세서

발명의 명칭: 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 장치 및 그 방법

기술분야

[1] 본 발명은 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 장치 및 그 방법에 관한 것으로, 보다 자세하게는 레이블 간 불균형이 존재하는 데이터 셋(dataset)에 대하여 전체 분류 성능을 향상시키기 위해 저빈도 레이블의 성능이 개선되도록 샘플 학습 데이터를 구성하는 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 장치 및 그 방법에 관한 것이다.

배경기술

[2] 인공 신경망을 기반으로 하는 딥러닝이 다양한 분야에서 기존의 기계 학습을 능가하는 우수한 성능을 보여주고 있으나, 신경망 모델이 특정 작업을 수행하기 위해서는 모델 가중치의 학습이 선행되어야 한다.

[3] 모델의 학습을 위해서는 학습 데이터가 필요하고 대부분의 분류를 위한 학습 데이터는 레이블 간의 불균형이 존재한다. 불균형이 존재하는 데이터를 사용하여 모델을 학습할 경우, 고빈도 레이블 샘플에 노이즈가 많거나 다른 레이블 샘플의 특징과 오버랩되는 등의 특수한 경우를 제외하고 저빈도 레이블의 성능이 떨어지기 때문에 레이블 샘플 수의 균형이 맞는 데이터를 사용한 모델보다 낮은 분류 성능이 나타난다.

[4] 이러한 레이블 불균형으로 인해 발생하는 학습 편향을 해결하기 위해 학습 데이터를 표본 추출하여 정보량을 조정하거나, 손실함수를 개선하여 샘플에 가중치를 부여하는 방법 등이 제안되어 왔다.

[5] 이와 관련하여, 한국공개특허 제2021-0120489호는 "레이블 데이터 생성 방법 및 이를 이용한 학습방법 및 장치"에 관하여 개시하고 있다.

발명의 상세한 설명

기술적 과제

[6] 본 발명은 상기와 같은 문제점을 해결하기 위해 발명된 것으로서, 전체 학습 데이터의 레이블 샘플수를 기반으로 하는 제1 샘플 추출기 및 전체 학습 데이터에 나타나는 정답 레이블 샘플수를 기반으로 하는 제2 샘플 추출기 중 하나를 이용하여 전체 학습 데이터에서 분류 모델의 학습에 사용할 샘플을 추출하여 샘플 학습 데이터를 구성하는 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 장치 및 그 방법을 제공하는데 그 목적이 있다.

[7] 또한, 본 발명은 레이블 샘플 수를 기반으로 구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습하고 업데이트하는 불균형 데이터에 대한 딥러닝

분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 장치 및 그 방법을 제공하는데 그 목적이 있다.

과제 해결 수단

[8] 상기의 목적을 달성하기 위한 본 발명에 따른 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 장치는 전체 학습 데이터의 레이블 샘플수를 기반으로 하는 제1 샘플 추출기 및 전체 학습 데이터에 나타나는 정답 레이블 샘플수를 기반으로 하는 제2 샘플 추출기 중 하나를 설정하는 샘플 추출기 설정부; 설정된 제1 샘플 추출기 또는 제2 샘플 추출기를 이용하여 전체 학습 데이터에서 분류 모델의 학습에 사용할 샘플을 추출하여 샘플 학습 데이터를 구성하는 학습 데이터 추출부; 및 구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습하고 업데이트하는 분류 모델 학습부;를 포함한다.

[9] 또한, 상기 샘플 추출기 설정부는 전체 학습 데이터에 동일한 추출 확률값을 할당한 후 비복원 추출을 수행하는 제1 샘플 추출기 및 전체 학습 데이터에 나타나는 정답 레이블 샘플을 산출하고, 산출된 정답 레이블 샘플에 동일한 추출 확률값을 할당한 후 복원 추출을 수행하는 제2 샘플 추출기 중 적어도 어느 하나로 설정되는 것을 특징으로 한다.

[10] 또한, 상기 샘플 추출기 설정부는 이후 구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습한 결과, 분류 모델의 검증 성능이 갱신되지 않은 것으로 판단되면, 설정된 샘플 추출기를 다른 샘플 추출기로 전환시키는 것을 특징으로 한다.

[11] 또한, 상기 분류 모델 학습부는 구성된 샘플 학습 데이터를 사용하여 딥러닝 기반의 분류 모델을 학습하고 업데이트되는 분류 모델의 가중치를 저장하는 것을 특징으로 한다.

[12] 또한, 구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습하고 업데이트한 결과에 대하여 분류 모델의 검증 성능 갱신 여부를 판단하는 분류 모델 검증 성능 판단부를 포함하는 것을 특징으로 한다.

[13] 상기의 목적을 달성하기 위한 본 발명에 따른 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 방법은 샘플 추출기 설정부에 의해, 전체 학습 데이터의 레이블 샘플수를 기반으로 하는 제1 샘플 추출기 및 전체 학습 데이터에 나타나는 정답 레이블 샘플수를 기반으로 하는 제2 샘플 추출기 중 하나를 설정하는 단계; 학습 데이터 추출부에 의해, 설정된 제1 샘플 추출기 또는 제2 샘플 추출기를 이용하여 전체 학습 데이터에서 분류 모델의 학습에 사용할 샘플을 추출하여 샘플 학습 데이터를 구성하는 단계; 및 분류 모델 학습부에 의해, 구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습하고 업데이트하는 단계;를 포함한다.

[14] 또한, 구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습하고

업데이트하는 단계는, 구성된 샘플 학습 데이터를 사용하여 딥러닝 기반의 분류 모델을 학습하고 업데이트되는 분류 모델의 가중치를 저장하는 것을 특징으로 한다.

[15] 또한, 구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습하고 업데이트하는 단계 이후에, 구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습하고 업데이트한 결과에 대하여 분류 모델의 검증 성능 갱신 여부를 판단하는 단계를 포함하는 것을 특징으로 한다.

[16] 또한, 구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습하고 업데이트한 결과에 대하여 분류 모델의 검증 성능 갱신 여부를 판단하는 단계에서, 분류 모델의 검증 성능이 갱신되지 않은 것으로 판단되면, 설정된 샘플 추출기를 다른 샘플 추출기로 전환시키는 단계를 포함하는 것을 특징으로 한다.

발명의 효과

[17] 본 발명에 따른 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 장치 및 그 방법은 전체 학습 데이터의 레이블 샘플수를 기반으로 하는 제1 샘플 추출기 및 전체 학습 데이터에 나타나는 정답 레이블 샘플수를 기반으로 하는 제2 샘플 추출기 중 하나를 이용하여 전체 학습 데이터에서 분류 모델의 학습에 사용할 샘플을 추출하여 샘플 학습 데이터를 구성하고, 레이블 샘플 수를 기반으로 구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습하고 업데이트함으로써, 저빈도 레이블 샘플의 추출 확률을 높여 저빈도 레이블에 대한 학습량을 늘리고 이로 인해 파생되는 학습 편향을 억제할 수 있는 효과가 있다.

도면의 간단한 설명

[18] 도 1은 본 발명에 따른 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 장치의 구성을 설명하기 위한 도면이다.

[19] 도 2는 본 발명에 따른 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 방법의 순서를 설명하기 위한 순서도이다.

발명의 실시를 위한 최선의 형태

[20] 본 발명은 다양한 변경을 가할 수 있고 여러 가지 실시예를 가질 수 있는 바, 특정 실시예들을 도면에 예시하고 상세하게 설명하고자 한다.

[21] 그러나 이는 본 발명을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다. 각 도면을 설명하면서 유사한 참조부호를 유사한 구성요소에 대해 사용하였다.

[22] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할

것이다. 반면에, 어떤 구성요소가 다른 구성요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는, 중간에 다른 구성요소가 존재하지 않는 것으로 이해되어야 할 것이다.

[23] 본 출원에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 출원에서, "포함하다" 또는 "가지다" 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 가능선을 미리 배제하지 않는 것으로 이해되어야 한다.

[24] 이하, 첨부한 도면들을 참조하여, 본 발명의 바람직한 실시예를 보다 상세하게 설명하고자 한다. 이하, 도면상의 동일한 구성요소에 대해서는 동일한 참조부호를 사용하고 동일한 구성요소에 대해서 중복된 설명은 생략한다.

[25] 도 1은 본 발명에 따른 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 장치의 구성을 설명하기 위한 도면이다.

[26] 도 1을 참조하여 설명하면, 본 발명에 따른 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 장치(100)는 크게 샘플 추출기 설정부(110), 학습 데이터 추출부(120), 분류 모델 학습부(130) 및 분류 모델 검증 성능 판단부(140)를 포함한다.

[27] 샘플 추출기 설정부(110)는 전체 학습 데이터의 레이블 샘플수를 기반으로 하는 제1 샘플 추출기 및 전체 학습 데이터에 나타나는 정답 레이블 샘플수를 기반으로 하는 제2 샘플 추출기 중 하나를 설정한다.

[28] 샘플 추출기 설정부(110)는 이후 구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습한 결과, 분류 모델의 검증 성능이 갱신되지 않은 것으로 판단되면, 설정된 샘플 추출기를 다른 샘플 추출기로 전환시킬 수 있다.

[29] 즉, 샘플 추출기 설정부(110)는 2가지 방식의 샘플 추출기를 가지며 현재 학습 경향에 따라 두 가지 샘플 추출기를 전환하여 이후 분류 모델의 학습 데이터를 추출에 사용되도록 한다.

[30] 여기서, 제1 샘플 추출기는 전체 학습 데이터에 동일한 추출 확률값을 할당한 후 비복원 추출을 수행하는 추출기이고, 제2 샘플 추출기는 전체 학습 데이터에 나타나는 정답 레이블 샘플을 산출하고, 산출된 정답 레이블 샘플에 동일한 추출 확률값을 할당한 후 복원 추출을 수행하는 추출기이다.

[31] 학습 데이터 추출부(120)는 설정된 제1 샘플 추출기 또는 제2 샘플 추출기를 이용하여 전체 학습 데이터에서 이후 분류 모델의 학습에 사용할 샘플을 추출하여 샘플 학습 데이터를 구성한다. 여기서, 입력되는 전체 학습 데이터는 분류 레이블이 부착된 모든 분야의 데이터로 딥러닝 기반 모델이 처리할 수 있는 형태로 정제된 상태이다.

[32] 학습 데이터 추출부(120)는 제1 샘플 추출기를 이용하는 경우, 전체 학습

데이터에 모두 동일한 추출 확률값을 할당하고 비복원 추출을 수행하여 샘플 학습 데이터를 구성한다.

[33] 학습 데이터 추출부(120)는 제2 샘플 추출기를 이용하는 경우, 전체 학습 데이터에 나타나는 정답 레이블 샘플 수를 산출하고, 산출된 정답 레이블 샘플에 동일한 추출 확률값을 할당한 후 복원 추출을 수행하여 샘플 학습 데이터를 구성한다. 보다 자세하게, 학습 데이터 추출부(120)는 제2 샘플 추출기를 이용하는 경우 정답 레이블 샘플 수를 토대로 저빈도 샘플의 추출 확률이 높아지는 것을 목적으로 하는 하기의 수식을 통해 샘플 각각에 대한 추출 확률을 계산할 수 있다.

[34] 즉, 제2 샘플 추출기는 레이블 샘플 수의 역수를 추출 확률로 정의한다. 입력 문장 x_i와 정답 레이블 y_i로 구성된 학습 데이터 $D = \{(x_i, y_i)\}_{i=1}^{N}$이 주어질 때, i 번째 샘플의 추출 확률 $P(x_i|D, y_i)$ 을 구하는 수식은 다음과 같다.

[35] [수식1]

$$N(a,b) = \begin{cases} 1, & if\ a = b \\ 0, & otherwise \end{cases}$$

$$P(x_i|D, y_i) = \frac{1}{\sum_j^D N(j, y_i)}$$

[36] 분류 모델 학습부(130)는 구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습하고 업데이트한다.

[37] 분류 모델 학습부(130)는 구성된 샘플 학습 데이터를 사용하여 딥러닝(Deep Learning) 기반의 분류 모델을 학습하고 업데이트되는 분류 모델의 가중치를 저장한다.

[38] 분류 모델 검증 성능 판단부(140)는 구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습하고 업데이트한 결과에 대하여 분류 모델의 검증 성능 갱신 여부를 판단한다.

발명의 실시를 위한 형태

[39] 도 2는 본 발명에 따른 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 방법의 순서를 설명하기 위한 순서도이다.

[40] 도 2를 참조하여 설명하면, 본 발명에 따른 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 방법은 앞서 설명한 본 발명에 따른 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 장치를 이용하는 것으로, 이하 중복되는 설명은 생략하기로 한다.

[41] 먼저, 전체 학습 데이터의 레이블 샘플수를 기반으로 하는 제1 샘플 추출기 및 전체 학습 데이터에 나타나는 정답 레이블 샘플수를 기반으로 하는 제2 샘플 추출기 중 하나를 설정한다(S100).

[42] S100 단계는 2가지 방식의 샘플 추출기를 가지며 현재 학습 경향에 따라 두 가지 샘플 추출기를 전환하여 이후 분류 모델의 학습 데이터를 추출에 사용되도록 한다.

[43] 다음, 설정된 제1 샘플 추출기 또는 제2 샘플 추출기를 이용하여 전체 학습 데이터에서 이후 분류 모델의 학습에 사용할 샘플을 추출하여 샘플 학습 데이터를 구성한다(S200). 여기서, 입력되는 전체 학습 데이터는 분류 레이블이 부착된 모든 분야의 데이터로 딥러닝 기반 모델이 처리할 수 있는 형태로 정제된 상태이다.

[44] S200 단계에서, 제1 샘플 추출기를 이용하는 경우, 전체 학습 데이터에 모두 동일한 추출 확률값을 할당하고 비복원 추출을 수행하여 샘플 학습 데이터를 구성한다.

[45] S200 단계에서, 제2 샘플 추출기를 이용하는 경우, 전체 학습 데이터에 나타나는 정답 레이블 샘플 수를 산출하고, 산출된 정답 레이블 샘플에 동일한 추출 확률값을 할당한 후 복원 추출을 수행하여 샘플 학습 데이터를 구성한다.

[46] 다음, 구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습하고 업데이트한다(S300).

[47] S300 단계는 구성된 샘플 학습 데이터를 사용하여 딥러닝(Deep Learning) 기반의 분류 모델을 학습하고 업데이트되는 분류 모델의 가중치를 저장한다.

[48] 다음, 구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습하고 업데이트한 결과에 대하여 분류 모델의 검증 성능 갱신 여부를 판단한다(S400).

[49] S400 단계에서, 분류 모델의 검증 성능이 갱신된 것으로 판단되면, S300 단계로 돌아간다.

[50] 한편, S400 단계에서, 분류 모델의 검증 성능이 갱신되지 않은 것으로 판단되면, 분류 모델의 검증 성능이 갱신 횟수와 사용자가 사전에 지정한 횟수를 비교한다(S500).

[51] S500 단계에서, 분류 모델의 검증 성능이 갱신 횟수와 사용자가 사전에 지정한 횟수만큼 갱신에 실패하지 않은 경우 설정된 샘플 추출기를 다른 샘플 추출기로 전환시키기 위해 S100 단계로 돌아간다.

[52] 한편, S500 단계에서, 분류 모델의 검증 성능이 갱신 횟수와 사용자가 사전에 지정한 횟수만큼 갱신에 실패한 경우 전체 학습을 중단하고 최종 모델을 저장한다.

[53]

[54] 이상 본 명세서에서 설명한 기능적 동작과 본 주제에 관한 실시형태들은 본 명세서에서 개시한 구조들 및 그들의 구조적인 등가물을 포함하여 디지털 전자 회로나 컴퓨터 소프트웨어, 펌웨어 또는 하드웨어에서 또는 이들 중 하나 이상이 조합에서 구현 가능하다.

[55] 본 명세서에서 기술하는 주제의 실시형태는 하나 이상이 컴퓨터 프로그램 제품, 다시 말해 데이터 처리 장치에 의한 실행을 위하여 또는 그 동작을

제어하기 위하여 유형의 프로그램 매체 상에 인코딩되는 컴퓨터 프로그램 명령에 관한 하나 이상이 모듈로서 구현될 수 있다. 유형의 프로그램 매체는 전파형 신호이거나 컴퓨터로 판독 가능한 매체일 수 있다. 전파형 신호는 컴퓨터에 의한 실행을 위하여 적절한 수신기 장치로 전송하기 위한 정보를 인코딩하기 위하여 생성되는 예컨대 기계가 생성한 전기적, 광학적 또는 전자기 신호와 같은 인공적으로 생성된 신호이다. 컴퓨터로 판독 가능한 매체는 기계로 판독 가능한 저장장치, 기계로 판독 가능한 저장 기판, 메모리 장치, 기계로 판독 가능한 전파형 신호에 영향을 미치는 물질의 조합 또는 이들 중 하나 이상이 조합일 수 있다.

[56] 부가적으로, 본 특허문헌에서 기술하는 논리 흐름과 구조적인 블록도는 개시된 구조적인 수단의 지원을 받는 대응하는 기능과 단계의 지원을 받는 대응하는 행위 및/또는 특정한 방법을 기술하는 것으로, 대응하는 소프트웨어 구조와 알고리즘과 그 등가물을 구축하는 데에도 사용 가능하다.

[57] 그리고 본 기술한 설명은 본 발명의 최상의 모드를 제시하고 있으며, 본 발명을 설명하기 위하여, 그리고 당업자가 본 발명을 제작 및 이용할 수 있도록 하기 위한 예를 제공하고 있다. 이렇게 작성된 명세서는 그 제시된 구체적인 용어에 본 발명을 제한하는 것이 아니다.

[58] 따라서, 상술한 예를 참조하여 본 발명을 상세하게 설명하였지만, 당업자라면 본 발명의 범위를 벗어나지 않으면서도 본 예들에 대한 개조, 변경 및 변형을 가할 수 있다. 요컨대 본 발명이 의도하는 효과를 달성하기 위해 도면에 도시된 모든 기능 블록을 별도로 포함하거나 도면에 도시된 모든 순서를 도시된 순서 그대로 따라야만 하는 것은 아니며, 그렇지 않더라도 얼마든지 청구항에 기재된 본 발명의 기술적 범위에 속할 수 있음에 주의한다.

산업상 이용가능성

[59] 인공신경망을 기반으로 하는 딥러닝 기술 분야에 이용할 수 있다.

청구범위

[청구항 1] 전체 학습 데이터에 동일한 추출 확률값을 할당한 후 비복원 추출을 수행하는 제1 샘플 추출기 및 전체 학습 데이터에 나타나는 정답 레이블 샘플을 산출하고, 산출된 정답 레이블 샘플별로 동일한 추출 확률값을 할당한 후 복원 추출을 수행하는 제2 샘플 추출기 중 하나를 설정하는 샘플 추출기 설정부;

설정된 제1 샘플 추출기 또는 제2 샘플 추출기를 이용하여 전체 학습 데이터에서 분류 모델의 학습에 사용할 샘플을 추출하여 샘플 학습 데이터를 구성하는 학습 데이터 추출부; 및

구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습하고 업데이트하는 분류 모델 학습부;를 포함하며,

상기 샘플 추출기 설정부는 이후 구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습한 결과, 분류 모델의 검증 성능이 갱신되지 않은 것으로 판단되면, 설정된 샘플 추출기를 다른 샘플 추출기로 전환시키는 것을 특징으로 하는 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 장치.

[청구항 2] 제1항에 있어서,

상기 분류 모델 학습부는 구성된 샘플 학습 데이터를 사용하여 딥러닝(Deep Learning) 기반의 분류 모델을 학습하고 업데이트되는 분류 모델의 가중치를 저장하는 것을 특징으로 하는 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 장치.

[청구항 3] 제1항에 있어서,

구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습하고 업데이트한 결과에 대하여 분류 모델의 검증 성능 갱신 여부를 판단하는 분류 모델 검증 성능 판단부를 포함하는 것을 특징으로 하는 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 장치.

[청구항 4] 샘플 추출기 설정부에 의해, 전체 학습 데이터에 동일한 추출 확률값을 할당한 후 비복원 추출을 수행하는 제1 샘플 추출기 및 전체 학습 데이터에 나타나는 정답 레이블 샘플을 산출하고, 산출된 정답 레이블 샘플별로 동일한 추출 확률값을 할당한 후 복원 추출을 수행하는 제2 샘플 추출기 중 하나를 설정하는 단계;

학습 데이터 추출부에 의해, 설정된 제1 샘플 추출기 또는 제2 샘플 추출기를 이용하여 전체 학습 데이터에서 분류 모델의 학습에 사용할 샘플을 추출하여 샘플 학습 데이터를 구성하는 단계;

분류 모델 학습부에 의해, 구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습하고 업데이트하는 단계; 및

구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습하고 업데이트한

결과에 대하여 분류 모델의 검증 성능 갱신 여부를 판단하는 단계를 포함하며,

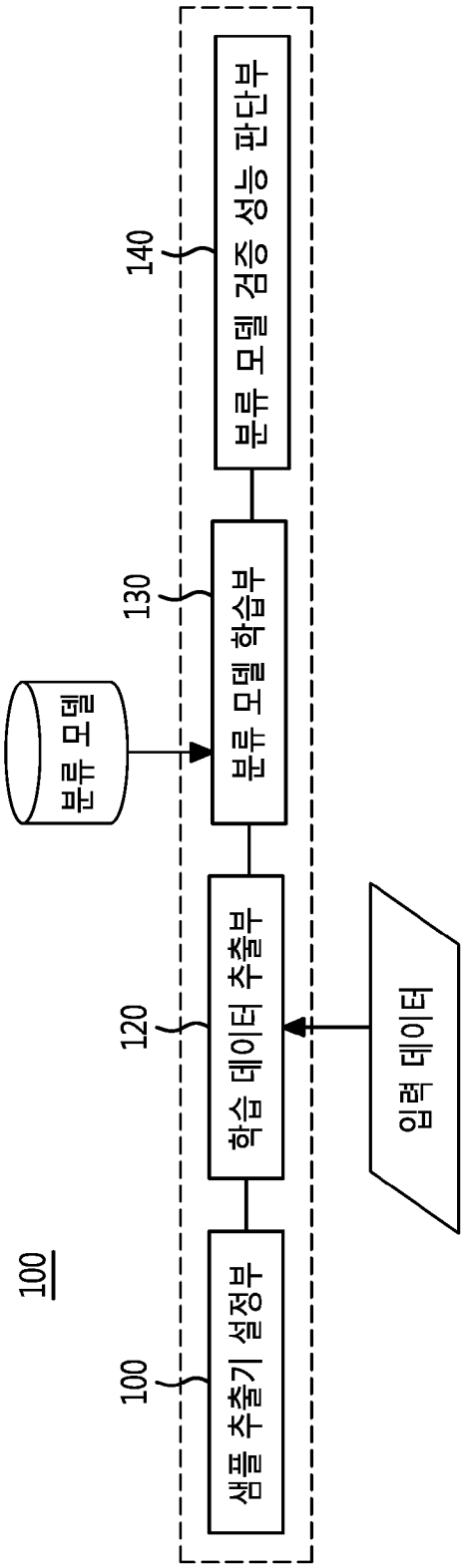
상기 분류 모델의 검증 성능 갱신 여부를 판단하는 단계에서, 분류 모델의 검증 성능이 갱신되지 않은 것으로 판단되면, 설정된 샘플 추출기를 다른 샘플 추출기로 전환시키는 단계를 포함하는 것을 특징으로 하는 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 방법.

[청구항 5] 제4항에 있어서,

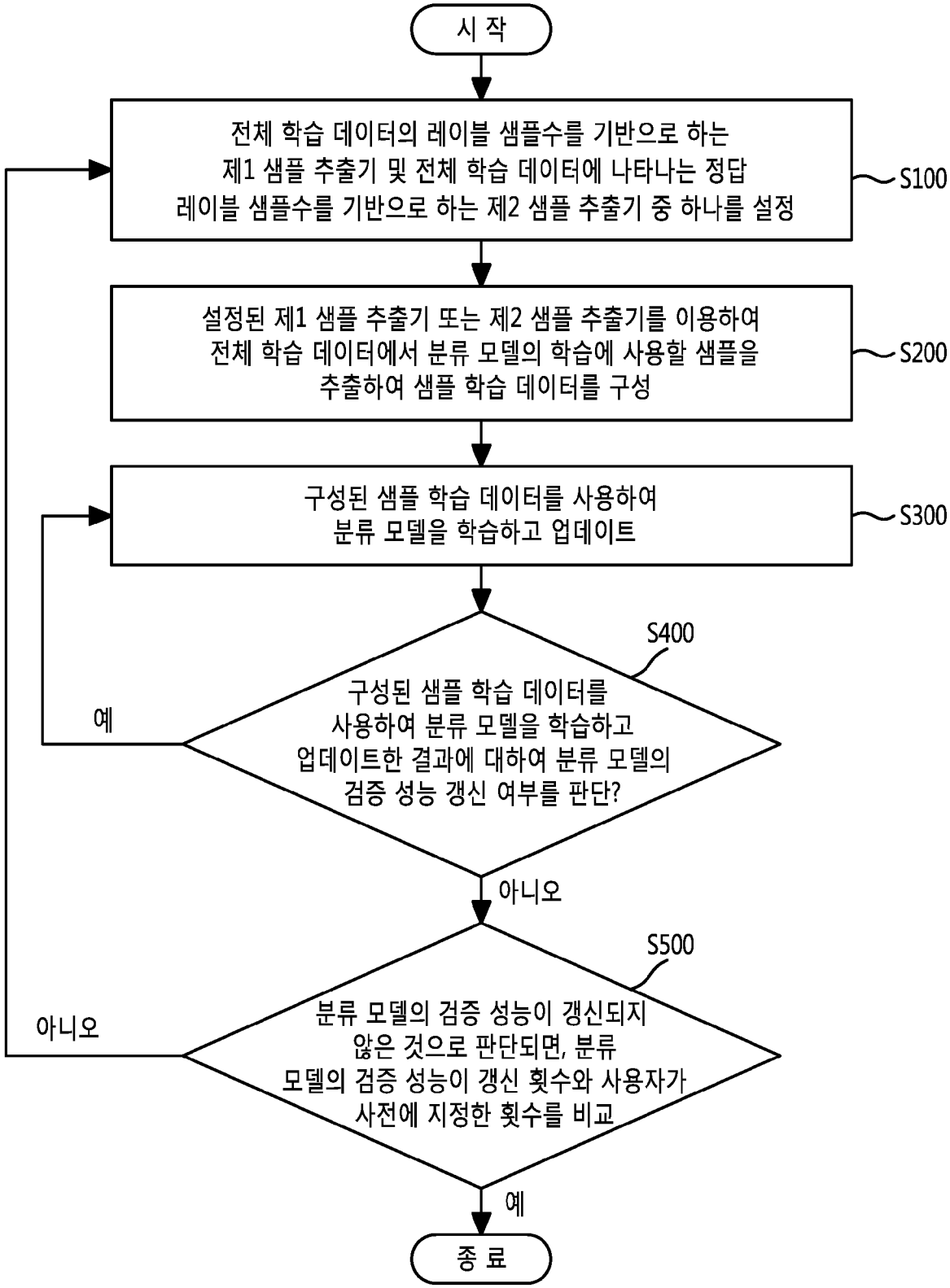
구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습하고 업데이트하는 단계는,

구성된 샘플 학습 데이터를 사용하여 딥러닝 기반의 분류 모델을 학습하고 업데이트되는 분류 모델의 가중치를 저장하는 것을 특징으로 하는 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 방법.

[도1]


100
샘플 추출기 설정부
100
학습 데이터 추출부
120
분류 모델 학습부
130
분류 모델 검증 성능 판단부
140
분류 모델
입력 데이터

[도2]


시 작
전체 학습 데이터의 레이블 샘플수를 기반으로 하는
제1 샘플 추출기 및 전체 학습 데이터에 나타나는 정답
레이블 샘플수를 기반으로 하는 제2 샘플 추출기 중 하나를 설정
S100
설정된 제1 샘플 추출기 또는 제2 샘플 추출기를 이용하여
전체 학습 데이터에서 분류 모델의 학습에 사용할 샘플을
추출하여 샘플 학습 데이터를 구성
S200
구성된 샘플 학습 데이터를 사용하여
분류 모델을 학습하고 업데이트
S300
S400
구성된 샘플 학습 데이터를
사용하여 분류 모델을 학습하고
업데이트한 결과에 대하여 분류 모델의
검증 성능 갱신 여부를 판단?
예
아니오
S500
분류 모델의 검증 성능이 갱신되지
않은 것으로 판단되면, 분류
모델의 검증 성능이 갱신 횟수와 사용자가
사전에 지정한 횟수를 비교
아니오
예
종 료

INTERNATIONAL SEARCH REPORT

International application No.
PCT/KR2022/019879

A. CLASSIFICATION OF SUBJECT MATTER

G06N 3/08(2006.01)i; **G06N 3/04**(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 3/08(2006.01); G06K 9/62(2006.01); G06K 9/72(2006.01); G06N 20/00(2019.01); G06N 3/04(2006.01); G10L 13/10(2013.01); G10L 19/038(2013.01)

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: 샘플(sample), 추출(extraction), 딥러닝(deep learning), 분류 모델(classification model), 레이블(label)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	SHIMIZU, Ryota et al. Balanced Mini-Batch Training for Imbalanced Image Data Classification with Neural Network. 2018 First International Conference on Artificial Intelligence for Industries. 26 September 2018. Retrieved from <https://ieeexplore.ieee.org/document/8665709>. See sections II and III and figure 1.	1-5
A	KR 10-2021-0085278 A (ELECTRONICS AND TELECOMMUNICATIONS RESEARCH INSTITUTE) 08 July 2021 (2021-07-08) See paragraphs [0029]-[0065] and figure 1.	1-5
A	KR 10-2263245 B1 (LG ELECTRONICS INC.) 14 June 2021 (2021-06-14) See paragraphs [0106]-[0141] and figure 4.	1-5
A	KR 10-2020-0052446 A (SAMSUNG SDS CO., LTD.) 15 May 2020 (2020-05-15) See paragraphs [0032]-[0052] and figure 2.	1-5

☑ 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	Date of mailing of the international search report
20 March 2023	**20 March 2023**

Name and mailing address of the ISA/KR	Authorized officer
Korean Intellectual Property Office **Government Complex-Daejeon Building 4, 189 Cheongsa-ro, Seo-gu, Daejeon 35208**	
Facsimile No. **+82-42-481-8578**	Telephone No.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2018-0075324 A1 (YAHOO JAPAN CORPORATION) 15 March 2018 (2018-03-15) See claims 1, 15 and 16.	1-5
PX	KR 10-2458360 B1 (CHANGWON NATIONAL UNIVERSITY INDUSTRY ACADEMY COOPERATION CORPS) 25 October 2022 (2022-10-25) See paragraphs [0026]-[0038] and figure 1. * This document is a published earlier application that serves as a basis for claiming priority of the present international application.	1-5
X	성수진 등. 불균형 범주 분류를 위한 동적 샘플링 스케줄러. 제33회 한글 및 한국어 정보처리 학술대회 논문집 (2021년). 15 October 2021 (SEONG, Su-Jin et al. Dynamic Sampling Scheduler for Unbalanced Data Classification. Proceedings of the 33rd Annual Conference on Human and Language Technology (2021).) See sections 2 and 5. *This document is a document declaring 'exceptions to lack of novelty' by the applicant.	1-5

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/KR2022/019879

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
KR 10-2021-0085278 A	08 July 2021	None	
KR 10-2263245 B1	14 June 2021	KR 10-2019-0096877 A	20 August 2019
		US 11056096 B2	06 July 2021
		US 2020-0005764 A1	02 January 2020
KR 10-2020-0052446 A	15 May 2020	US 2020-0134454 A1	30 April 2020
US 2018-0075324 A1	15 March 2018	JP 2018-045361 A	22 March 2018
		JP 6199461 B1	20 September 2017
KR 10-2458360 B1	25 October 2022	None	

국 제 조 사 보 고 서

국제출원번호
PCT/KR2022/019879

A. 발명이 속하는 기술분류(국제특허분류(IPC))

G06N 3/08(2006.01)i; **G06N 3/04**(2006.01)i

B. 조사된 분야

조사된 최소문헌(국제특허분류를 기재)

G06N 3/08(2006.01); G06K 9/62(2006.01); G06K 9/72(2006.01); G06N 20/00(2019.01); G06N 3/04(2006.01); G10L 13/10(2013.01); G10L 19/038(2013.01)

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: 샘플(sample), 추출(extraction), 딥러닝(deep learning), 분류 모델(classification model), 레이블(label)

C. 관련 문헌

카테고리*	인용문헌명 및 관련 구절(해당하는 경우)의 기재	관련 청구항
A	RYOTA SHIMIZU 등. "Balanced Mini-Batch Training for Imbalanced Image Data Classification with Neural Network". 2018 First International Conference on Artificial Intelligence for Industries, 2018.09.26. <https://ieeexplore.ieee.org/document/8665709> 섹션 II, III 및 도면 1	1-5
A	KR 10-2021-0085278 A (한국전자통신연구원) 2021.07.08 단락 [0029]-[0065] 및 도면 1	1-5
A	KR 10-2263245 B1 (엘지전자 주식회사) 2021.06.14 단락 [0106]-[0141] 및 도면 4	1-5
A	KR 10-2020-0052446 A (삼성에스디에스 주식회사) 2020.05.15 단락 [0032]-[0052] 및 도면 2	1-5
A	US 2018-0075324 A1 (YAHOO JAPAN CORPORATION) 2018.03.15 청구항 1, 15, 16	1-5

☑ 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	국제조사보고서 발송일
2023년03월20일(20.03.2023)	**2023년03월20일(20.03.2023)**

ISA/KR의 명칭 및 우편주소	심사관
대한민국 특허청 **(35208) 대전광역시 서구 청사로 189, 4동 (둔산동, 정부대전청사)**	**양정록**
팩스 번호 **+82-42-481-8578**	전화번호 **+82-42-481-5709**

서식 PCT/ISA/210 (두 번째 용지) (2022년 7월)

국 제 조 사 보 고 서

국제출원번호
PCT/KR2022/019879

C. 관련 문헌

카테고리*	인용문헌명 및 관련 구절(해당하는 경우)의 기재	관련 청구항
PX	KR 10-2458360 B1 (창원대학교 산학협력단) 2022.10.25 단락 [0026]-[0038] 및 도면 1 * 위 문헌은 본 국제출원의 우선권주장의 기초가 되는 선출원의 공개된 공보임	1-5
X	성수진 등. "불균형 범주 분류를 위한 동적 샘플링 스케줄러", 제33회 한글 및 한국어 정보처리 학술대회 논문집(2021년), 2021.10.15. 섹션 2, 5 *본 문헌은 출원인에 의해 '신규성 상실의 예외'가 선언된 문헌임.	1-5

국 제 조 사 보 고 서
대응특허에 관한 정보

국제출원번호
PCT/KR2022/019879

국제조사보고서에서 인용된 특허문헌	공개일	대응특허문헌	공개일
KR 10-2021-0085278 A	2021/07/08	없음	
KR 10-2263245 B1	2021/06/14	KR 10-2019-0096877 A	2019/08/20
		US 11056096 B2	2021/07/06
		US 2020-0005764 A1	2020/01/02
KR 10-2020-0052446 A	2020/05/15	US 2020-0134454 A1	2020/04/30
US 2018-0075324 A1	2018/03/15	JP 2018-045361 A	2018/03/22
		JP 6199461 B1	2017/09/20
KR 10-2458360 B1	2022/10/25	없음	