(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2023년05월11일
(11) 등록번호 10-2531758
(24) 등록일자 2023년05월08일

(51) 국제특허분류(Int. Cl.)
G06N 20/00 (2019.01) *G06F 18/00* (2023.01)
(52) CPC특허분류
G06N 20/00 (2021.08)
G06F 18/24 (2023.01)
(21) 출원번호 10-2021-0020521
(22) 출원일자 2021년02월16일
심사청구일자 2021년02월16일
(65) 공개번호 10-2022-0116984
(43) 공개일자 2022년08월23일
(56) 선행기술조사문헌
황선희 외. Exploiting Transferable Knowledge for Fairness-aware Image Classification. ACCV(2020). 2020.12.*
(뒷면에 계속)

(73) 특허권자
연세대학교 산학협력단
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
(72) 발명자
변혜란
서울특별시 서대문구 연세로 50, D810호 (신촌동)
황선희
서울특별시 서대문구 연세로 50, D810호 (신촌동)
(뒷면에 계속)
(74) 대리인
특허법인(유한)아이시스

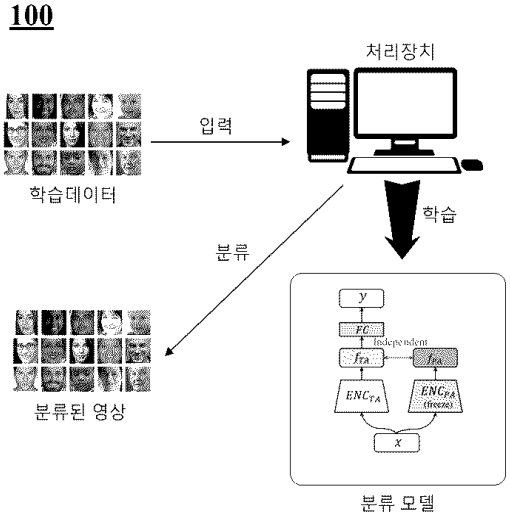
전체 청구항 수 : 총 6 항 심사관 : 박성수

(54) 발명의 명칭 **공정한 영상 분류를 위한 학습 방법 및 영상을 공정하게 분류하는 장치**

(57) 요 약

개시된 기술은 공정한 영상 분류를 위한 학습 방법 및 영상을 공정하게 분류하는 장치에 관한 것으로, 프로세서가 보호속성 분류기에 학습데이터를 입력하여 상기 학습데이터에 포함된 객체의 분류와 관련된 보호속성 정보를 토대로 상기 보호속성 분류기를 학습시키는 단계; 및 상기 프로세서가 상기 학습된 보호속성 분류기의 인코더를 이용하여 획득된 상기 보호속성 정보와 연관성이 없는 속성 정보를 이용하여 타겟속성 분류기를 학습시키는 단계;를 포함한다.

대 표 도 - 도1



(52) CPC특허분류
G06V 10/469 (2023.01)
G06V 40/16 (2022.01)

(72) 발명자
박성호
서울특별시 서대문구 연세로 50, D810호 (신촌동)
이필현
서울특별시 서대문구 연세로 50, D810호 (신촌동)
전석규
서울특별시 서대문구 연세로 50, D810호 (신촌동)
김도형
서울특별시 서대문구 연세로 50, D810호 (신촌동)

(56) 선행기술조사문헌
Yaroslav Ganin et al. Domain-Adversarial Training of Neural Networks. arXiv:1505.07818v4 [stat.ML]. 2016.5.26.
T Wang et al. Balanced Datasets Are Not Enough:Estimating and Mitigating Gender Bias in Deep Image Representations. arXiv:1811.08489v4 [cs.CV]. 2019.10.*
김병주 외. Learning Not to Learn: Training Deep Neural Networks with Biased Data. arXiv:1812.10352v2 [cs.CV]. 2019.04.15.
Brian Hu Zhang et al. Mitigating Unwanted Biases with Adversarial Learning. arXiv:1801.07593v1 [cs.LG]. 2018.1.22.
*는 심사관에 의하여 인용된 문헌

이 발명을 지원한 국가연구개발사업

과제고유번호	1711102850
과제번호	2019-0-01396-002
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신방송연구개발사업
연구과제명	인공지능 모델과 학습데이터의 편향성 분석-탐지-완화·제거 지원 프레임워크 개발
기 여 율	1/1
과제수행기관명	한국과학기술원
연구기간	2020.01.01 ~ 2020.12.31

공지예외적용 : 있음

명 세 서

청구범위

청구항 1

프로세서가 보호속성 분류기에 학습데이터를 입력하여 상기 학습데이터에 포함된 객체의 분류와 관련된 보호속성 정보를 토대로 상기 보호속성 분류기를 학습시키는 단계; 및

상기 프로세서가 상기 학습된 보호속성 분류기의 인코더를 이용하여 획득된 상기 보호속성 정보와 연관성이 없는 속성 정보를 이용하여 타겟속성 분류기를 학습시키는 단계;를 포함하되,

상기 타겟속성 분류기는 크로스 엔트로피 손실이 최소화 되도록 학습하고, 상기 보호속성 분류기에서 유클리디안 거리가 먼 것으로 분류된 특징벡터를 거리가 가까운 것으로 학습하고, 상기 보호속성 분류기에서 유클리디안 거리가 가까운 것으로 분류된 특징벡터를 거리가 먼 것으로 학습되고, 동시에 상기 학습데이터에 포함된 복수개의 영상들 각각에 대한 특징을 추출하고, 유사한 특징을 갖는 샘플들을 그룹으로 클러스터링하고, 서로 다른 특징을 갖는 샘플들은 Wasserstein 거리가 가까워지도록 학습하고, 클러스터링 된 그룹 내 서브그룹을 형성하여 상기 서브그룹들의 wasserstein 거리가 가까워지도록 학습되는, 공정한 영상 분류를 위한 학습 방법.

청구항 2

제 1 항에 있어서,

상기 보호속성 정보는 상기 학습데이터에 포함된 복수의 인물들 각각에 대한 나이, 인종 및 성별의 라벨값을 포함하는 공정한 영상 분류를 위한 학습 방법.

청구항 3

제 1 항에 있어서,

상기 보호속성 분류기는 크로스 엔트로피 손실을 최소화하도록 학습되는 공정한 영상 분류를 위한 학습 방법.

청구항 4

삭제

청구항 5

삭제

청구항 6

삭제

청구항 7

얼굴 영상을 입력받는 입력장치;

보호속성 분류기 및 타겟속성 분류기를 포함하는 분류 모델을 저장하는 저장장치; 및

상기 타겟속성 분류기를 이용하여 상기 얼굴 영상의 특징을 분류하는 처리장치;를 포함하되,

상기 보호속성 분류기는 학습데이터에 포함된 객체의 분류와 관련된 보호속성 정보를 토대로 학습되고, 상기 타겟속성 분류기는 상기 학습된 보호속성 분류기의 인코더를 이용하여 획득된 보호속성 정보와 연관성이 없는 속성 정보를 이용하여 사전에 학습되고,

상기 타겟속성 분류기는 크로스 엔트로피 손실이 최소화 되도록 학습하고, 상기 보호속성 분류기에서 유클리디안 거리가 먼 것으로 분류된 특징벡터를 거리가 가까운 것으로 학습하고, 상기 보호속성 분류기에서 유클리디안 거리가 가까운 것으로 분류된 특징벡터를 거리가 먼 것으로 학습되고, 동시에 상기 학습데이터에 포함된 복수개의 영상들 각각에 대한 특징을 추출하고, 유사한 특징을 갖는 샘플들을 그룹으로 클러스터링하고, 서로 다른 특

징을 갖는 샘플들은 Wasserstein 거리가 가까워지도록 학습하고, 클러스터링 된 그룹 내 서브그룹을 형성하여 상기 서브그룹들의 wasserstein 거리가 가까워지도록 학습되는, 영상을 공정하게 분류하는 장치.

청구항 8

제 7 항에 있어서,

상기 보호속성 정보는 상기 학습데이터에 포함된 복수의 인물들 각각에 대한 나이, 인종 및 성별의 라벨값을 포함하는 영상을 공정하게 분류하는 장치.

청구항 9

제 7 항에 있어서,

상기 보호속성 분류기는 크로스 엔트로피 손실을 최소화하도록 학습되는 영상을 공정하게 분류하는 장치.

청구항 10

삭제

청구항 11

삭제

청구항 12

삭제

발명의 설명

기 술 분 야

[0001] 개시된 기술은 공정하게 영상을 분류하기 위한 학습 방법 및 영상을 공정하게 분류하는 장치에 관한 것이다.

배 경 기 술

[0002] 종래 머신러닝 기반 영상 분류 기법은 시스템에서 의도하지 않게 편향된 분류 결과를 나타낸다. 예컨대, 여성의 영상이 입력될 경우 쇼핑이나 요리와 같은 캡션으로 편향될 수 있고 남성의 영상이 입력될 경우 운전이나 촬영과 같은 캡션으로 편향될 수 있다. 또한, 사람의 얼굴을 인식하는 시스템인 경우 백인의 얼굴은 제대로 인식하지만 흑인의 얼굴은 잘 인식하지 못하는 경우도 발생할 수 있다.

[0003] 이와 같이 영상을 분류할 때 편향된 결과가 발생하지 않도록 하기 위해서 종래에는 도메인 적응 기법이나 적대적 편향성 제거 또는 풀린 표현 학습(Disentangled Representation Learning)과 같은 방법을 이용하여 특징 표현에서 보호된 속성과 관련된 정보를 제외하는 기법을 이용하였다. 이러한 기법에 따라 영상 분류에 대한 공정성은 향상될 수 있으나 보호된 속성의 추가적인 주석을 획득하기 위해 많은 시간이 소모되며 실제 시스템에 적용 시 실행이 불가능하거나 모델의 확장성을 제한하는 문제가 있었다.

선행기술문헌

특허문헌

[0004] (특허문헌 0001) 한국 등록특허 제10-2116396호

발명의 내용

해결하려는 과제

[0005] 개시된 기술은 공정하게 영상을 분류하기 위한 학습 방법 및 영상을 공정하게 분류하는 장치를 제공하는데 있다.

과제의 해결 수단

[0006] 상기의 기술적 과제를 이루기 위하여 개시된 기술의 제 1 측면은 프로세서가 보호속성 분류기에 학습데이터를 입력하여 상기 학습데이터에 포함된 객체의 분류와 관련된 보호속성 정보를 토대로 상기 보호속성 분류기를 학습시키는 단계 및 상기 프로세서가 상기 학습된 보호속성 분류기의 인코더를 이용하여 획득된 상기 보호속성 정보와 연관성이 없는 속성 정보를 이용하여 타겟속성 분류기를 학습시키는 단계를 포함하는 공정한 영상 분류를 위한 학습 방법을 제공하는데 있다.

[0007] 상기의 기술적 과제를 이루기 위하여 개시된 기술의 제 2 측면은 얼굴 영상을 입력받는 입력장치, 보호속성 분류기 및 타겟속성 분류기를 포함하는 분류 모델을 저장하는 저장장치 및 상기 타겟속성 분류기를 이용하여 상기 얼굴 영상의 특징을 분류하는 처리장치를 포함하되, 상기 보호속성 분류기는 학습데이터에 포함된 객체의 분류와 관련된 보호속성 정보를 토대로 학습되고, 상기 타겟속성 분류기는 상기 학습된 보호속성 분류기의 인코더를 이용하여 획득된 보호속성 정보와 연관성이 없는 속성 정보를 이용하여 사전에 학습되는 영상을 공정하게 분류하는 장치를 제공하는데 있다.

발명의 효과

[0008] 개시된 기술의 실시 예들은 다음의 장점들을 포함하는 효과를 가질 수 있다. 다만, 개시된 기술의 실시 예들이 이를 전부 포함하여야 한다는 의미는 아니므로, 개시된 기술의 권리범위는 이에 의하여 제한되는 것으로 이해되어서는 아니 될 것이다.

[0009] 개시된 기술의 일 실시예에 따르면 공정하게 영상을 분류하기 위한 학습 방법 및 영상을 공정하게 분류하는 장치는 영상 분류에 따른 편향성을 방지하는 효과가 있다.

[0010] 또한, 보호속성 정보를 직접적으로 사용하지 않고 두 분류기가 서로 독립된 정보를 학습하게 하여 높은 공정성을 나타내는 효과가 있다.

도면의 간단한 설명

[0011] 도 1은 개시된 기술의 일 실시예에 따른 공정하게 영상을 분류하기 위한 학습 과정을 나타낸 도면이다.

도 2는 개시된 기술의 일 실시예에 따른 공정하게 영상을 분류하기 위한 학습 방법에 대한 순서도이다.

도 3은 개시된 기술의 일 실시예에 따른 영상을 분류하는 장치에 대한 블록도이다.

도 4는 개시된 기술의 일 실시예에 따른 영상 분류 모델의 구조를 나타낸 도면이다.

도 5는 보호속성 분류기를 학습하는 과정을 나타낸 도면이다.

도 6은 타겟속성 분류기를 학습하는 과정을 나타낸 도면이다.

도 7은 독립된 특징의 삼중항 손실(Feature Independency Triplet Loss)을 이용하여 학습하는 과정을 나타낸 도면이다.

도 8은 그룹 차원의 공정 손실(Group-wise Fair Loss)을 이용하여 학습하는 과정을 나타낸 도면이다.

발명을 실시하기 위한 구체적인 내용

[0012] 본 발명은 다양한 변경을 가할 수 있고 여러 가지 실시예를 가질 수 있는 바, 특정 실시예들을 도면에 예시하고 상세한 설명에 상세하게 설명하고자 한다. 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.

[0013] 제 1 , 제 2, A, B 등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 해당 구성요소들은 상기 용어들에 의해 한정되지는 않으며, 단지 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다. 예를 들어, 본 발명의 권리 범위를 벗어나지 않으면서 제 1 구성요소는 제 2 구성요소로 명명될 수 있고, 유사하게 제 2 구성요소도 제 1 구성요소로 명명될 수 있다. 및/또는 이라는 용어는 복수의 관련된 기재된 항목들의 조합 또는 복수의 관련된 기재된 항목들 중의 어느 항목을 포함한다.

[0014] 본 명세서에서 사용되는 용어에서 단수의 표현은 문맥상 명백하게 다르게 해석되지 않는 한 복수의 표현을 포함

하는 것으로 이해되어야 한다. 그리고 "포함한다" 등의 용어는 설시된 특징, 개수, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함을 의미하는 것이지, 하나 또는 그 이상의 다른 특징들이나 개수, 단계 동작 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 배제하지 않는 것으로 이해되어야 한다.

[0015] 도면에 대한 상세한 설명을 하기에 앞서, 본 명세서에서의 구성부들에 대한 구분은 각 구성부가 담당하는 주기능 별로 구분한 것에 불과함을 명확히 하고자 한다. 즉, 이하에서 설명할 2개 이상의 구성부가 하나의 구성부로 합쳐지거나 또는 하나의 구성부가 보다 세분화된 기능별로 2개 이상으로 분화되어 구비될 수도 있다.

[0016] 그리고 이하에서 설명할 구성부 각각은 자신이 담당하는 주기능 이외에도 다른 구성부가 담당하는 기능 중 일부 또는 전부의 기능을 추가적으로 수행할 수도 있으며, 구성부 각각이 담당하는 주기능 중 일부 기능이 다른 구성부에 의해 전담되어 수행될 수도 있음은 물론이다. 따라서, 본 명세서를 통해 설명되는 각 구성부들의 존재 여부는 기능적으로 해석되어야 할 것이다.

[0017] 도 1은 개시된 기술의 일 실시예에 따른 공정하게 영상을 분류하기 위한 학습 과정을 나타낸 도면이다. 도 1을 참조하면 본 발명에 따른 학습장치는 영상을 공정하게 분류하기 위한 분류 모델을 탑재한다. 그리고 분류 모델은 2개의 분류기를 포함한다.

[0018] 처리장치는 영상을 분류하기 위해서 사전에 학습데이터를 이용하여 분류 모델을 학습시킨다. 처리장치는 분류 모델의 영상 분류 결과가 편향되는 것을 방지하기 위해서 두 분류기에 각각 서로 독립된 정보를 학습시켜서 공정하게 영상을 분류할 수 있는 학습 과정을 수행한다.

[0019] 한편, 분류 모델이 포함하는 2개의 분류기는 각각 보호속성 분류기와 타겟속성 분류기이다. 보호속성 분류기는 입력되는 학습데이터의 보호속성 정보를 이용하여 학습데이터의 보호 속성을 분류하도록 학습된다. 보호속성 정보는 학습데이터에 포함된 객체를 분류하기 위한 정보를 의미한다. 그리고 타겟속성 분류기는 보호속성 분류기와 서로 상반되는 정보를 학습하여 이후 영상이 입력됐을 때 분류하는 과정을 수행하게 된다. 즉, 보호속성 분류기는 타겟속성 분류기를 학습시키기 위해서 이용되는 것이며 학습이 완료된 이후에는 타겟속성 분류기에 영상을 입력하여 분류하는 것이다.

[0020] 한편, 분류 모델을 학습하는데 이용하는 학습데이터는 보호속성에 대한 정보를 라벨값으로 포함하는 영상을 의미한다. 예컨대, 학습데이터는 각 인물들에 대한 나이, 인종 및 성별을 보호속성 정보로 포함하는 복수의 인물들에 대한 얼굴 영상일 수 있다. 일반적으로 다수의 영상을 하나의 데이터셋 형태로 분류 모델에 입력될 수 있다. 학습데이터 또한 다수의 영상을 포함하는 데이터셋일 수 있다. 이하에서는 서로 다른 복수의 인물들에 대한 영상을 학습데이터로 가정한다.

[0021] 한편, 처리장치는 도 1에 따른 학습 과정을 수행하기 위해서 학습데이터를 먼저 보호속성 분류기에 입력한다. 보호속성 분류기는 학습데이터에 포함된 보호속성 정보를 이용하여 학습을 수행한다. 보호속성 분류기는 학습데이터가 입력되면 나이, 인종 및 성별 정보를 분류하게 되는데, 이때 크로스 엔트로피 손실(Cross Entropy Loss)을 최소화하는 방향으로 학습을 수행할 수 있다. 크로스 엔트로피는 실제 분포값을 알지 못하는 상태에서, 모델링을 통해 구한 값을 이용하여 실제 분포값을 예측하는 것이다. 일 실시예로, 보호속성 분류기가 학습데이터에 대한 나이, 인종 및 성별 정보를 분류한 결과를 이용하여 보호속성 정보를 예측하는 것일 수 있다. 일반적으로 크로스 엔트로피에서는 실제값과 예측값이 맞는 경우에는 0으로 수렴하고, 실제값과 예측값이 서로 다른 경우에는 값이 커지기 때문에, 학습 과정에서 실제값과 예측값의 차이를 줄이기 위한 손실 함수를 구축할 수 있다.

[0022] 한편, 이와 같이 보호속성 분류기가 학습되면, 타겟속성 분류기도 보호속성 분류기와 동일한 학습데이터를 입력받아서 학습을 수행한다. 이때, 타겟속성 분류기는 보호속성 분류기의 인코더가 분류한 특징과 상반되는 특징을 이용하여 학습될 수 있다. 즉, 보호속성 분류기가 나이, 인종 및 성별을 구분하도록 학습되는 것과는 반대로 타겟속성 분류기는 나이, 인종 및 성별을 구분할 수 없도록 학습될 수 있다. 타겟속성 분류기는 보호속성 분류기와 마찬가지로 크로스 엔트로피 손실을 최소화 하는 방향으로 학습을 수행할 수 있다. 이때, 크로스 엔트로피 손실과 함께 독립된 특징의 삼중항 손실(Feature Independency Triplet Loss) 및 그룹 차원의 공정 손실(Group-wise Fair Loss)을 이용하여 학습을 수행할 수 있다.

[0023] 한편, 타겟속성 분류기에서 이용하는 Feature Independency Triplet Loss는 타겟속성 분류기가 보호속성 분류기와 서로 상반되는 특징을 학습하도록 하는 손실값이다. 예컨대, 보호속성 분류기에서 유클리디안 거리가 먼 것으로 분류된 특징벡터를 타겟속성 분류기에서 거리가 가까운 것으로 학습할 수 있다. 마찬가지로 보호속성 분류기에서 유클리디안 거리가 가까운 것으로 분류된 특징벡터는 타겟속성 분류기에서는 거리가 먼 것으로 학습할

수 있다. 이러한 학습 과정에 따라 타겟속성 분류기는 학습데이터에 포함된 인물의 나이, 인종 및 성별을 구분하지 못하도록 학습될 수 있다.

[0024] 한편, 타겟속성 분류기는 Group-wise Fair Loss에 따라 아래의 3 단계로 학습을 수행할 수 있다. 먼저 학습데이터를 인코더에 입력하여 학습데이터에 포함된 복수개의 영상들 각각에 대한 특징을 추출하고, 유사한 특징들끼리 그룹으로 클러스터링할 수 있다. 입력된 학습데이터는 복수의 영상들을 포함하는 데이터셋이므로 유사한 특징들을 그룹으로 묶을 수 있다. 두 번째로, 서로 다른 특징들은 Wasserstein 거리가 가까워지도록 학습할 수 있다. 이는 보호속성 정보가 다른 두 그룹에 대하여 구분이 안되는 특징을 학습하기 위한 과정을 의미한다. 세 번째로, 클러스터링 된 그룹 내 서브그룹을 형성하여 서브그룹들의 Wasserstein 거리가 가까워지도록 학습할 수 있다.

[0025] 상술한 바와 같이 타겟속성 분류기는 보호속성 분류기와 서로 상반되는 특징을 학습할 수 있으며 크로스 엔트로피 손실, Feature Independency Triplet Loss 및 Group-wise Fair Loss의 3가지 손실을 합산한 결과가 최소화되도록 학습할 수 있다.

[0026] 도 2는 개시된 기술의 일 실시예에 따른 공정하게 영상을 분류하기 위한 학습 방법에 대한 순서도이다. 도 2를 참조하면 공정한 영상 분류를 위한 학습 방법(200)은 보호속성 분류기를 학습시키는 단계(210) 및 타겟속성 분류기를 학습시키는 단계(220)를 포함한다. 학습 방법(200)은 영상을 분류하기 위한 장치의 프로세서를 통해 수행될 수 있으며 210 단계 및 220 단계가 순차적으로 수행될 수 있다.

[0027] 210 단계에서 프로세서는 보호속성 분류기에 학습데이터를 입력하고 학습데이터에 대한 보호속성 정보를 이용하여 보호속성 분류기가 보호 속성을 분류하도록 학습시킨다. 프로세서는 2개의 분류기 중 보호속성 분류기에 먼저 학습데이터를 입력할 수 있다. 보호속성 분류기는 보호속성 정보를 이용하여 학습데이터에 포함된 인물의 나이, 인종 및 성별을 분류할 수 있다.

[0028] 220 단계에서 프로세서는 학습된 보호속성 분류기의 인코더를 이용하여 보호속성 분류기가 학습한 특징과 서로 상반되는 특징을 학습하도록 타겟속성 분류기를 학습시킨다. 프로세서는 보호속성 분류기의 인코더를 이용하여 획득된 보호속성 정보와 연관성이 없는 속성 정보를 이용하여 타겟속성 분류기를 학습시킴으로써 두 분류기가 서로 독립된 정보로 학습되도록 제어할 수 있다. 보호속성 분류기를 학습시킨 이후에 프로세서는 보호속성 분류기에 입력한 학습데이터와 동일한 데이터를 타겟속성 분류기에도 입력할 수 있다. 이때, 타겟속성 분류기는 인물의 나이, 인종 및 성별을 구분하지 못하는 방향으로 학습된다. 이를 위해서 타겟속성 분류기는 크로스 엔트로피 손실, Feature Independency Triplet Loss 및 Group-wise Fair Loss을 이용하여 학습을 수행한다.

[0029] 도 3은 개시된 기술의 일 실시예에 따른 공정하게 영상을 분류하기 위한 학습 장치에 대한 블록도이다. 도 3을 참조하면 영상을 공정하게 분류하는 장치(300)는 입력장치(310), 저장장치(320) 및 처리장치(330)를 포함한다.

[0030] 입력장치(310)는 얼굴 영상을 입력받는다. 입력장치는 사용자가 얼굴 영상을 분류 장치(330)에 입력하기 위해서 이용할 수 있는 인터페이스 장치일 수 있다. 예컨대, 키보드나 마우스와 같이 데이터를 입력할 수 있는 장치일 수 있다.

[0031] 저장장치(320)는 분류 모델을 저장한다. 분류 모델은 보호속성 분류기 및 타겟속성 분류기를 포함한다. 저장장치(320)는 일정 이상의 용량을 가진 분류 모델을 저장할 수 있고 필요 시 로딩할 수 있다. 예컨대, 하드디스크나 SSD와 같은 메모리를 이용할 수 있다.

[0032] 처리장치(330)는 타겟속성 분류기를 이용하여 입력된 얼굴 영상의 특징을 분류한다. 얼굴 영상을 분류하기 이전에 처리장치는 분류 모델을 학습시킬 수 있다. 예컨대,보호속성 분류기에 학습데이터를 입력하고 학습데이터에 포함된 객체의 분류와 관련된 보호속성 정보를 이용하여 보호속성 분류기를 학습시킬 수 있다. 그리고 학습된 보호속성 분류기의 인코더를 통해 획득된 보호속성 정보와 연관성이 없는 속성 정보를 이용하여 타겟속성 분류기를 학습시킬 수 있다. 이와 같은 학습 과정을 거친 이후 얼굴 영상이 입력되면 공정한 기준에 따라 영상의 특징을 분류하게 된다. 처리장치(330)는 분류 장치의 CPU 또는 AP와 같이 데이터를 처리할 수 있는 장치로 구현될 수 있다. 처리장치(330)는 앞서 도 1을 통해 설명한 바와 같이 두 분류기를 학습시킬 수 있다.

[0033] 한편, 상술한 바와 같은 분류 장치(300)는 컴퓨터에서 실행될 수 있는 실행가능한 알고리즘을 포함하는 프로그램(또는 어플리케이션)으로 구현될 수도 있다. 상기 프로그램은 일시적 또는 비일시적 판독 가능 매체(non-transitory computer readable medium)에 저장되어 제공될 수 있다.

[0034] 비일시적 판독 가능 매체란 레지스터, 캐쉬, 메모리 등과 같이 짧은 순간 동안 데이터를 저장하는 매체가 아니

라 반영구적으로 데이터를 저장하며, 기기에 의해 판독(reading)이 가능한 매체를 의미한다. 구체적으로는, 상술한 다양한 어플리케이션 또는 프로그램들은 CD, DVD, 하드 디스크, 블루레이 디스크, USB, 메모리카드, ROM (read-only memory), PROM (programmable read only memory), EPROM(Erasable PROM, EPROM) 또는 EEPROM(Electrically EPROM) 또는 플래시 메모리 등과 같은 비일시적 판독 가능 매체에 저장되어 제공될 수 있다.

[0035] 일시적 판독 가능 매체는 스태틱 램(Static RAM, SRAM), 다이내믹 램(Dynamic RAM, DRAM), 싱크로너스 디램(Synchronous DRAM, SDRAM), 2배속 SDRAM(Double Data Rate SDRAM, DDR SDRAM), 증강형 SDRAM(Enhanced SDRAM, ESDRAM), 동기화 DRAM(Synclink DRAM, SLDRAM) 및 직접 램버스 램(Direct Rambus RAM, DRRAM) 과 같은 다양한 RAM을 의미한다.

[0036] 도 4는 개시된 기술의 일 실시예에 따른 영상 분류 모델의 구조를 나타낸 도면이다. 도 4를 참조하면 분류 모델은 보호속성 분류기 및 타겟속성 분류기를 포함한다. 보호속성 분류기의 인코더(401)과 타겟속성 분류기의 인코더(402)에는 동일한 학습데이터가 입력된다. 다만 보호속성 분류기에서는 학습데이터에 포함된 보호속성 정보를 라벨값으로 이용하여 학습을 수행하기 때문에 학습데이터에 포함된 인물의 나이, 인종 및 성별을 구분하도록 학습되는 반면, 타겟속성 분류기는 인물의 나이, 인종 및 성별을 구분하지 못하는 방향으로 학습될 수 있다. 즉, 두 분류기는 서로 독립된 정보를 이용하여 학습되며 이러한 과정에 따라 영상을 분류할 때 편향되지 않은 결과를 나타낼 수 있다. 이와 같은 학습이 수행된 이후에 인물의 영상이 입력되면 처리장치는 타겟속성 분류기만을 이용하여 영상을 분류할 수 있다. 즉, 보호속성 분류기는 타겟속성 분류기를 학습시키기 위해서만 이용된다.

[0037] 도 5는 보호속성 분류기를 학습하는 과정을 나타낸 도면이다. 도 5를 참조하면 보호속성 분류기는 학습데이터가 입력되면 보호속성에 대한 표현을 인코딩할 수 있다. 보호속성 분류기는 복수의 컨볼루션 레이어(Convolution layer)와 완전히 연결된(Fully Connected) 레이어들로 구성된다. 보호속성 분류기는 학습을 위한 데이터셋이 주어지면 나이, 인종 및 성별 각각에 대한 라벨값을 이용하여 최적화하는 과정을 수행할 수 있다. 즉, 크로스 엔트로피 손실이 최소화되도록 학습을 수행할 수 있다. 보호속성 분류기는 아래의 수학식 1에 따라 크로스 엔트로피 손실을 계산할 수 있다.

[0038] [수학식 1]

$$\mathcal{L}_{PAC} = -\sum_{i=1}^{n} g_i log(\hat{g}_i) - \sum_{i=1}^{n} a_i log(\hat{a}_i) - \sum_{i=1}^{n} r_i log(\hat{r}_i),$$

[0039]

[0041] 여기에서 $\hat{g}$ 는 성별(Gender)을 의미하고, $\hat{a}$ 는 나이(Age)를 의미하고, $\hat{r}$ 은 인종(Race)를 의미한다. 그리고 n 은 데이터의 개수를 의미한다.

[0042] 도 6은 타겟속성 분류기를 학습하는 과정을 나타낸 도면이다. 도 6을 참조하면 타겟속성 분류기의 구조 또한 보호속성 분류기와 마찬가지로 컨볼루션 레이어(Convolution layer)들과 완전히 연결된(Fully Connected) 레이어들로 구성된다. 학습데이터가 주어지면 타겟속성 분류기는 아래의 수학식 2에 따라 크로스 엔트로피 손실을 계산한다.

[0043] [수학식 2]

$$\mathcal{L}_{target} = -\sum_{i=1}^{m} t_i log(\hat{t}_i),$$

[0044]

[0046] 여기에서 $\hat{t}_i$ 는 각 속성별 분류기의 예측을 나타내고 m 는 대상 데이터셋의 샘플 수를 의미한다.

[0047] 한편, 타겟속성 분류기는 크로스 엔트로피 손실 뿐만 아니라 Feature Independency Triplet Loss 및 Group-wise Fair Loss를 이용하여 학습을 수행할 수 있다. 이하의 도 7 및 도 8을 통해 Feature Independency Triplet Loss 및 Group-wise Fair Loss를 계산하는 과정을 설명한다.

[0048] 도 7은 Feature Independency Triplet Loss를 이용하여 학습하는 과정을 나타낸 도면이다. 도 7을 참조하면 타겟속성 분류기는 Feature Independency Triplet Loss에 따라 보호속성 분류기의 인코더와 서로 독립된 학습을

수행할 수 있다. 일 실시예로, 도 7과 같이 데이터셋의 미니 배치 $X = [x_1, x_2, ..., x_k]$ 에서 각 앵커 샘플 x_a 에 대해 2개의 샘플 x_i 및 x_j 를 무작위로 선택할 수 있다. 여기에서 k 는 배치 사이즈를 의미한다.

[0049] 앵커 및 선택된 샘플을 사전에 훈련된 보호속성 분류기의 인코더에 의해 각각 인코딩된다. 즉, x_a 는 f_a 로 인코딩되고, x_i 는 f_i 로 인코딩되고, x_j 는 f_j 로 인코딩될 수 있다. 그리고 앵커의 특성인 f_a 를 기준으로 유클리드 거리 $d(f_a, f_i)$ 와 $d(f_a, f_j)$ 를 계산할 수 있다. 그리고 x_a 에서 더 먼 샘플을 양의 샘플 x_p 에 할당하고 다른 하나는 음의 샘플인 x_n 에 할당한다. 그리고 튜플 $[(x_a^1, x_p^1, x_n^1), (x_a^2, x_p^2, x_n^2), ..., (x_a^k, x_p^k, x_n^k)]$ 을 구성한다. 여기에서 음의 샘플 x_n^i 는 보호된 속성의 관점에서 앵커 샘플 x_a^i 와 더 유사할 수 있다. Feature Independency Triplet Loss는 이하의 수학식 3에 따라 정의될 수 있다.

[0050] [수학식 3]

$$\mathcal{L}_{triplet} = \sum_{i=1}^{N} \max(d(h_a^i, h_p^i) - d(h_a^i, h_n^i) + \alpha, 0),$$

[0051]

[0053] 여기에서 h_a^i , h_p^i 및 h_n^i 는 각각 x_a^i , x_p^i 및 x_n^i 에서 인코딩된 특징을 의미한다.

[0054] 도 8은 Group-wise Fair Loss를 이용하여 학습하는 과정을 나타낸 도면이다. 도 8을 참조하면 타겟속성 분류기는 Group-wise Fair Loss에 따라 모델의 공정성을 더욱 강화할 수 있다. Group-wise Fair Loss는 서로 다른 보호된 속성 그룹 간의 오분류 비율에 대한 불일치를 최소화하는 것을 목표로 하는 손실함수이다. 타겟속성 분류기는 Group-wise Fair Loss에 따라 학습데이터에 포함된 복수개의 영상들 각각에 대한 특징을 추출하고, 유사한 특징끼리 그룹으로 클러스터링할 수 있다. 아래 수학식 4와 같이 클러스터링을 수행할 수 있다.

[0055] [수학식 4]

$$minimize|P(\hat{y} \neq y|G_1) - P(\hat{y} \neq y|G_2)|,$$

[0056]

[0058] 여기에서 $\hat{y}$ 및 y 는 각각 예측 라벨값과 대상 라벨값을 나타낸다. 그리고 G1과 G2는 각각 다른 속성 그룹을 의미한다.

[0059] 한편, 이와 같이 두 그룹을 클러스터링하면 아래 수학식 5에 따라 두 그룹 간의 Wasserstein 거리를 계산할 수 있다.

[0060] [수학식 5]

$$\mathcal{W}(H_{G_1}, H_{G_2}) = \inf_{\gamma \in \prod(H_{G_1}, H_{G_2})} \mathbb{E}_{(z_1, z_2) \sim \gamma}[\|z_1 - z_2\|].$$

[0061]

[0063] 여기에서 보호된 속성 라벨값은 대상 데이터셋에 사용할 수 없기 때문에 보호된 속성 측면에서 그룹을 분리하기 위해 보호속성 인코더에서 전송된 보호된 속성의 정보를 이용한다. 즉, 수학식 5에 도시된 바와 같이 사전에 학습된 보호속성 분류기의 인코더를 이용하여 데이터셋 X_t 에서 특징 F 을 추출할 수 있다. 그리고 추출된 F 를 기반으로 K-평균 클러스터링 알고리즘을 사용하여 X_t 를 두 그룹 G1과 G2로 클러스터링할 수 있다. 그리고 G1과 G2 간의 Wasserstein 거리를 최소화할 수 있다.

[0064] 한편, 수학식 5에서 $\prod(H_{G_1}, H_{G_2})$ 는 한계 값이 각각 H_{G_1} 와 H_{G_2} 를 모두 연결하는 결합분포 $\gamma(z_1, z_2)$ 의 집합을 의미한다. Group-wise Fair Loss는 보호된 속성 그룹 간의 공정성을 개선하지만 대상의 속성 측면에서는 여전히 편향성이 존재한다. 따라서, 아래의 수학식 6과 같이 서브그룹의 특성 간 Wasserstein 거리를 최소화하

는 연산을 수행하여 편향성을 줄이는 방향으로 학습될 수 있다.

[0065] [수학식 6]

$$\mathcal{W}(H_{G_1^+}, H_{G_1^-}) = \inf_{\gamma \in \prod(H_{G_1^+}, H_{G_1^-})} \mathbb{E}_{(z_1,z_2)\sim\gamma}[\|z_1 - z_2\|].$$

$$\mathcal{W}(H_{G_2^+}, H_{G_2^-}) = \inf_{\gamma \in \prod(H_{G_2^+}, H_{G_2^-})} \mathbb{E}_{(z_1,z_2)\sim\gamma}[\|z_1 - z_2\|],$$

[0066]

[0068] 한편, 타겟속성 분류기는 아래 수학식 7에 따라 앞서 계산한 크로스 엔트로피 손실, 독립된 특징의 삼중항 손실(Feature Independency Triplet Loss) 및 그룹 차원의 공정 손실(Group-wise Fair Loss)을 합산한 결과가 최소값을 갖도록 학습될 수 있다.

[0069] [수학식 7]

$$\mathcal{L}_{TAC} = \lambda_1 \mathcal{L}_{target} + \lambda_2 \mathcal{L}_{triplet} + \lambda_3 \mathcal{L}_{group},$$

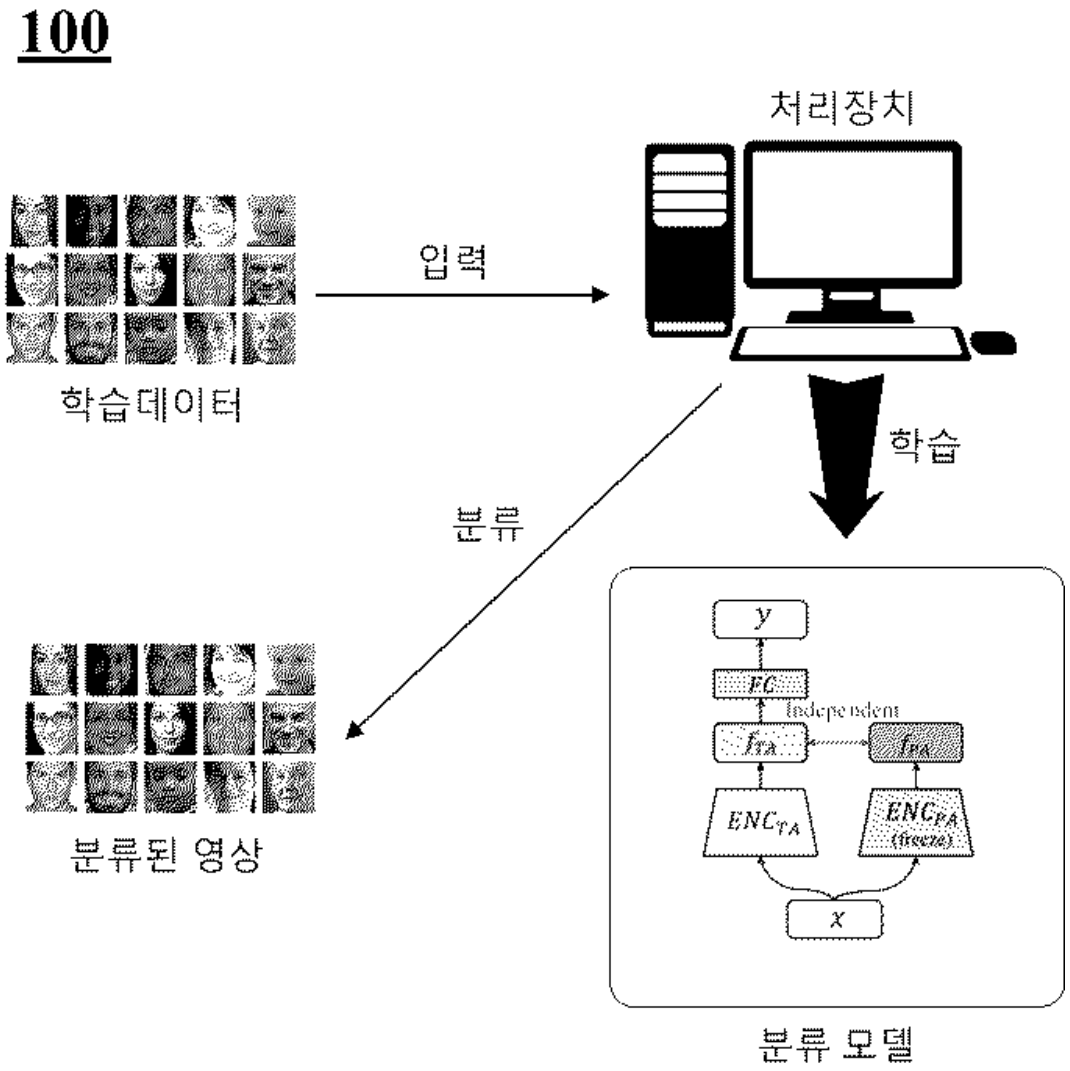
[0070]

[0072] 여기에서 λ_* 는 손실 균형을 맞추기 위한 하이퍼 파라미터를 의미한다. 이러한 과정에 따라 타겟속성 분류기는 보호속성 정보가 다른 두 그룹에 대하여 구분하지 못하도록 학습되기 때문에 영상 분류 시 편향되지 않은 분류 결과를 나타낼 수 있다.

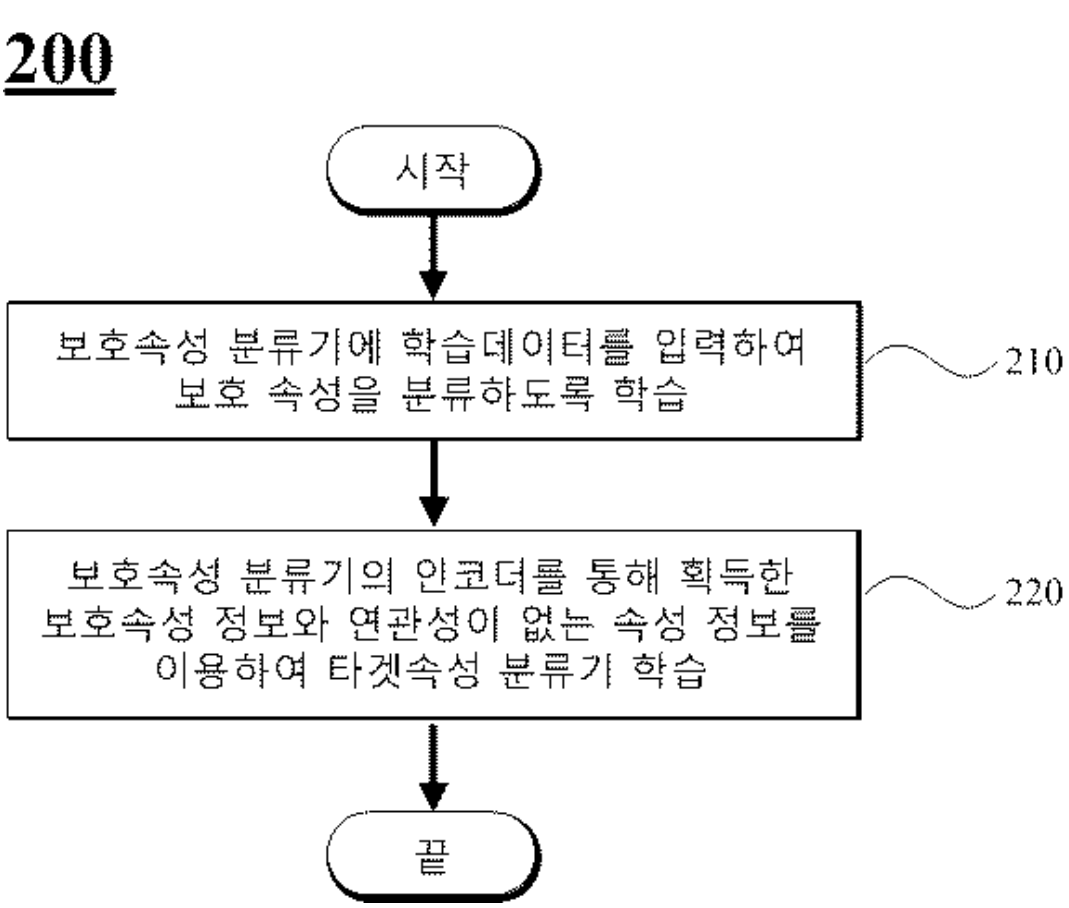
[0073] 개시된 기술의 일 실시예에 따른 공정한 영상 분류를 위한 학습 방법 및 영상을 공정하게 분류하는 장치는 이해를 돕기 위하여 도면에 도시된 실시 예를 참고로 설명되었으나, 이는 예시적인 것에 불과하며, 당해 분야에서 통상적 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시 예가 가능하다는 점을 이해할 것이다. 따라서, 개시된 기술의 진정한 기술적 보호범위는 첨부된 특허청구범위에 의해 정해져야 할 것이다.

도면

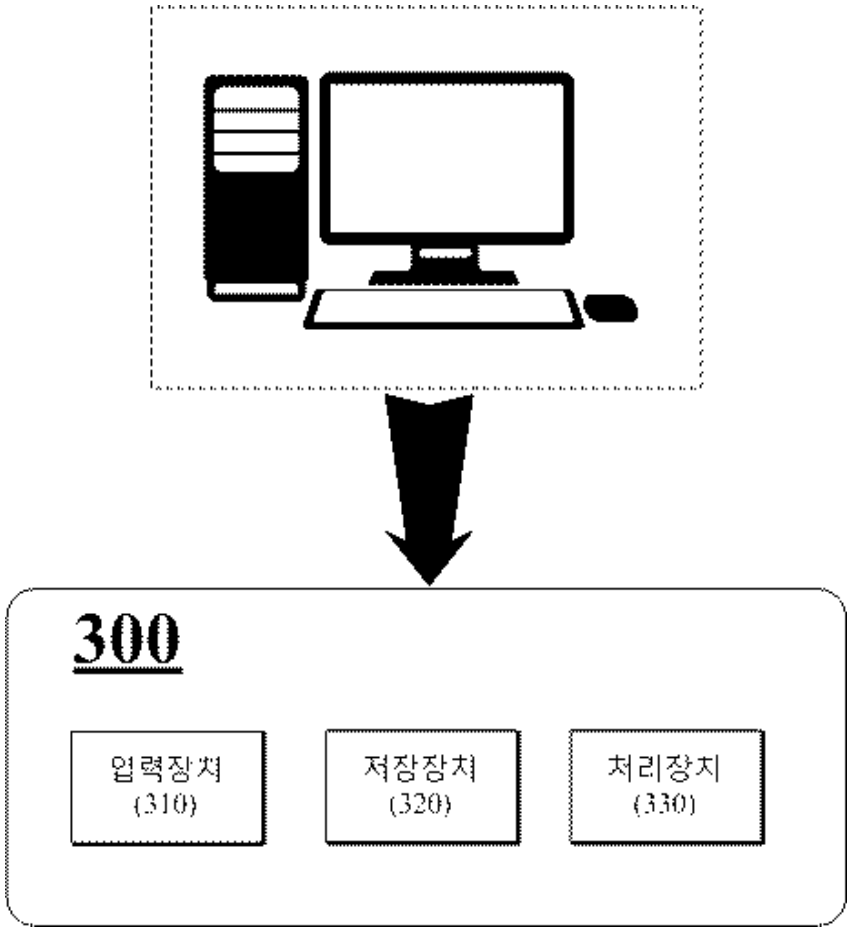
도면1


100
처리장치
입력
학습데이터
학습
분류
분류된 영상
y
FC
fc dependent
f_{TA}
f_{PA}
ENC_{TA}
ENC_{PA}
(freeze)
x
분류 모델

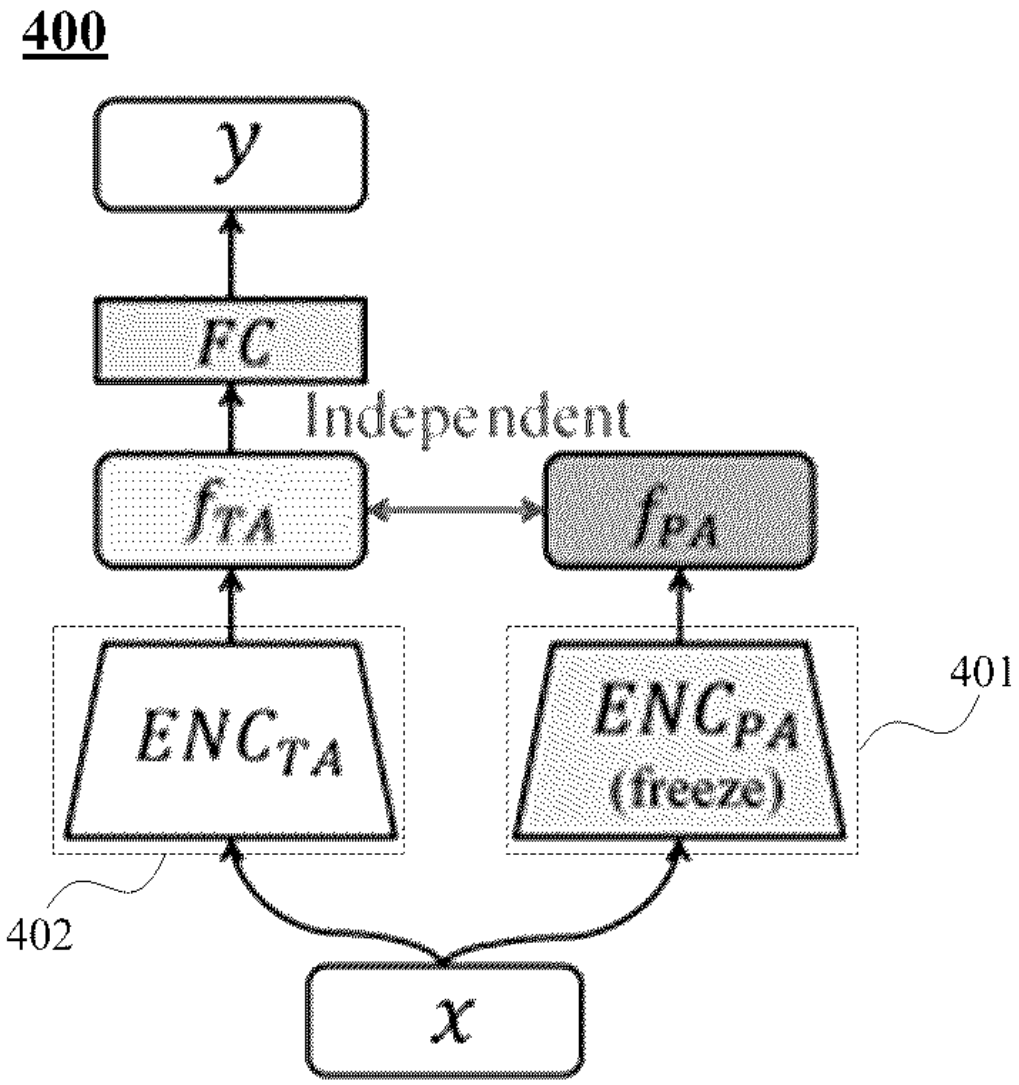
도면2


200
시작
보호속성 분류기에 학습데이터를 입력하여 보호 속성을 분류하도록 학습
210
보호속성 분류기의 인코더를 통해 획득한 보호속성 정보와 연관성이 없는 속성 정보를 이용하여 타겟속성 분류기 학습
220
끝

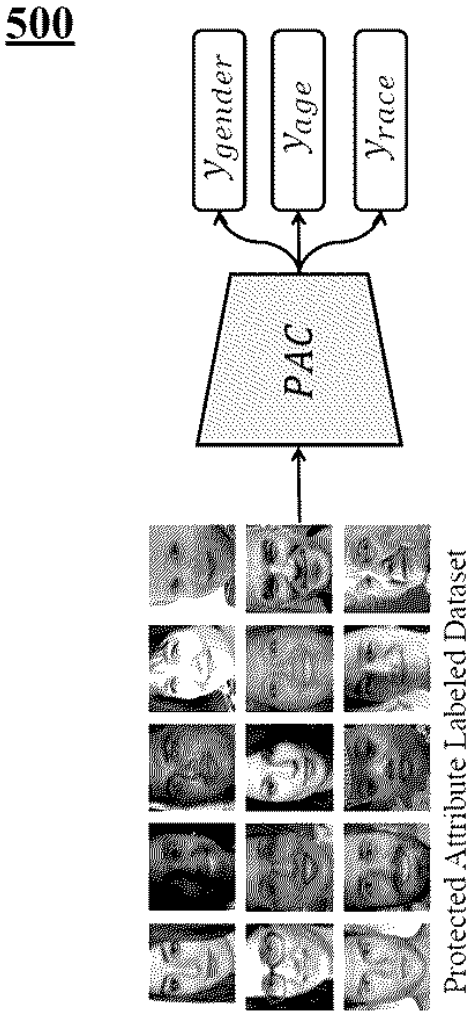
도면3



도면4

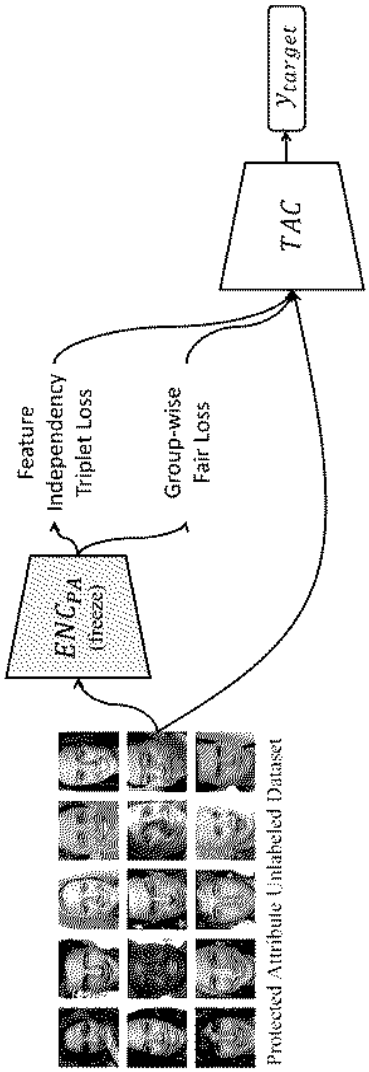


도면5


500
y_{gender}
y_{age}
y_{race}
PAC
Protected Attribute Labeled Dataset

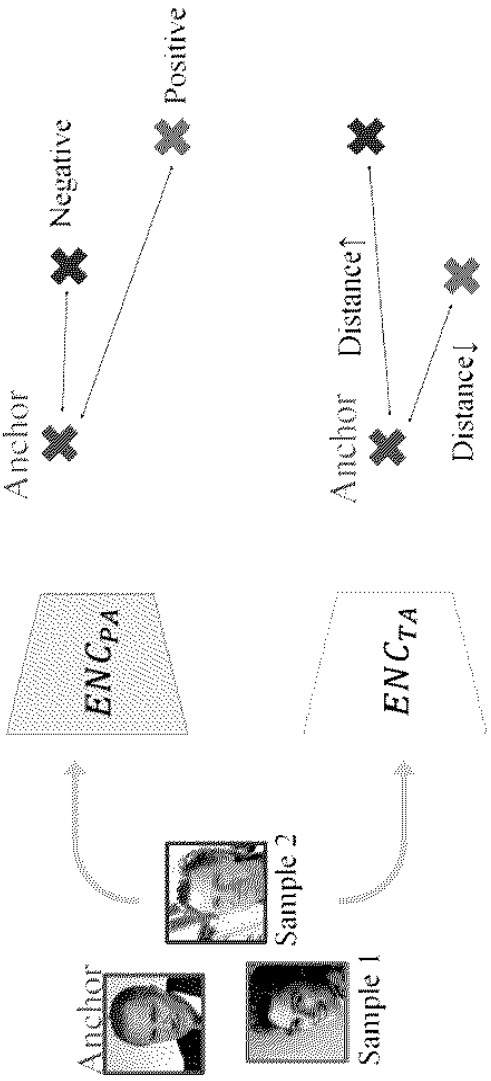
도면6

600

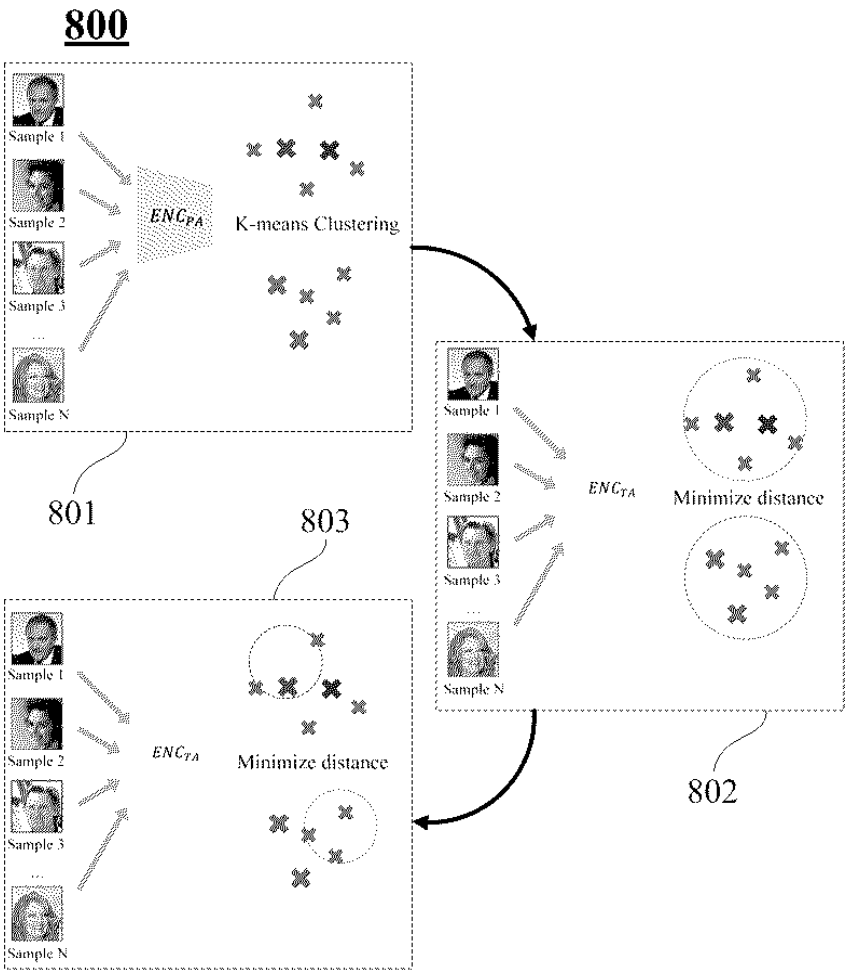

y_{target}
TAC
Feature Independency Triplet Loss
Group-wise Fair Loss
ENC_{PA} (freeze)
Protected Attribute Unlabeled Dataset

도면7

700


Anchor
Negative
Positive
Anchor
Distance↑
Distance↓
ENC_{PA}
ENC_{TA}
Anchor
Sample 1
Sample 2

도면8


800
Sample 1
Sample 2
Sample 3
...
Sample N
ENC_{PA}
K-means Clustering
801
803
ENC_{TA}
Minimize distance
802