(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0139590
(43) 공개일자 2022년10월17일

(51) 국제특허분류(Int. Cl.)
G06K 9/62 (2022.01) *G06N 20/00* (2019.01)
G06N 5/02 (2006.01) *G06T 7/60* (2017.01)
G06V 10/40 (2022.01)
(52) CPC특허분류
G06K 9/6267 (2022.01)
G06N 20/00 (2021.08)
(21) 출원번호 10-2021-0045735
(22) 출원일자 2021년04월08일
심사청구일자 2021년04월08일

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
석흥일
서울특별시 성북구 길음로 118, 411동 1202호(길음동, 길음뉴타운)
오관석
서울특별시 동대문구 약령시로 8, 교우빌딩 602호(제기동)
(74) 대리인
송인호, 윤형근, 최관락

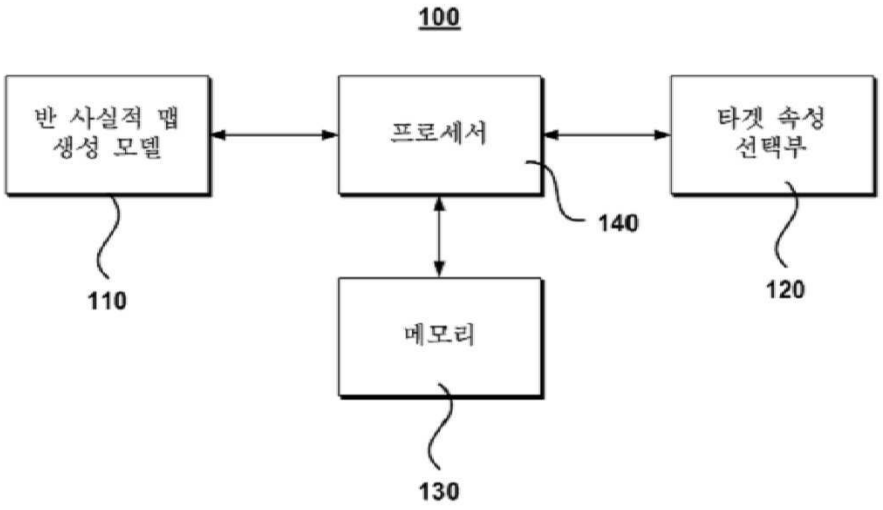
전체 청구항 수 : 총 7 항

(54) 발명의 명칭 **분류 결과 설명이 가능한 반 사실적 맵 생성 방법 및 그 장치**

(57) 요 약

분류 결과 설명이 가능한 반 사실적 맵 생성 방법 및 그 장치가 개시된다. 분류 결과 설명이 가능한 반 사실적 맵 생성 장치는, 입력 이미지와 타겟 레이블을 신경망 기반 반 사실적 맵 생성 모델에 적용하여 반 사실적 맵(counterfactual map)을 생성하는 단계; 상기 입력 이미지와 상기 반 사실적 맵간의 선형 합을 통해 합성된 타겟 이미지를 생성하는 단계; 및 상기 합성된 타겟 이미지의 특징값을 추출한 후 분류 모델에 의해 타겟 레이블로 분류되도록 학습하는 단계를 포함한다.

대 표 도 - 도1



(52) CPC특허분류
G06N 5/025 (2019.01)
G06T 7/60 (2013.01)
G06V 10/40 (2022.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711126387
과제번호 2017-0-01779-005
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 혁신성장동력프로젝트(R&D)(과기정통부)
연구과제명 의사결정 이유를 설명할 수 있는 인간 수준의 학습·추론 프레임워크 개발
기 여 율 2/3
과제수행기관명 한국과학기술원
연구기간 2021.01.01 ~ 2021.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1711126110
과제번호 2019-0-00079-003
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 정보통신방송혁신인재양성(R&D)
연구과제명 인공지능대학원지원(고려대학교)
기 여 율 1/3
과제수행기관명 고려대학교산학협력단
연구기간 2021.01.01 ~ 2021.12.31

명 세 서

청구범위

청구항 1

입력 이미지와 타겟 레이블을 신경망 기반 반 사실적 맵 생성 모델에 적용하여 반 사실적 맵(counterfactual map)을 생성하는 단계;

상기 입력 이미지와 상기 반 사실적 맵간의 선형 합을 통해 합성된 타겟 이미지를 생성하는 단계; 및

상기 합성된 타겟 이미지의 특징값을 추출한 후 분류 모델에 의해 타겟 레이블로 분류되도록 학습하는 단계를 포함하는 분류 결과 설명이 가능한 반 사실적 맵 생성 방법.

청구항 2

제1 항에 있어서,

상기 반 사실적 맵 생성 모델은, 상기 입력 이미지의 구조적 정보를 유지하도록 주기 일관성 손실을 고려하여 상기 입력 이미지에 의존적인 반 사실적 맵을 생성하는 것을 특징으로 하는 분류 결과 설명이 가능한 반 사실적 맵 생성 방법.

청구항 3

제1 항에 있어서,

상기 반 사실적 맵 생성 모델은 적대적 생성 신경망 기반 모델이되, 상기 합성된 타겟 이미지가 실제 샘플에 매핑되도록 판별부를 적대적 학습시키고, 상기 합성된 타겟 이미지가 타겟 레이블로 예측되기 위한 반 사실적 맵을 생성하도록 학습되는 분류 결과 설명이 가능한 반 사실적 맵 생성 방법.

청구항 4

제1 항에 있어서,

상기 타겟 레이블은 입력 이미지를 타겟으로 변환하기 위한 속성이 부여된 레이블인 것을 특징으로 하는 분류 결과 설명이 가능한 반 사실적 맵 생성 방법.

청구항 5

입력 이미지를 분석하여 특징값을 추출하는 특징 추출부;

상기 입력 이미지의 특징값들과 타겟 레이블을 신경망 기반 반 사실적 맵 생성 모델에 적용하여 반 사실적 맵(counterfactual map)을 생성하는 맵 생성부; 및

상기 입력 이미지와 상기 반 사실적 맵간의 선형 합을 통해 합성된 타겟 이미지를 생성하는 타겟 이미지 생성부를 포함하는 장치.

청구항 6

제5 항에 있어서,

상기 반 사실적 맵 생성 모델은 적대적 생성 신경망 기반 모델이되, 상기 합성된 타겟 이미지가 실제 샘플에 매

핑되도록 판별부를 적대적 학습하며, 상기 합성된 타겟 이미지가 타겟 레이블로 예측되기 위한 반 사실적 맵을 생성하도록 학습되는 것을 특징으로 하는 장치.

청구항 7

제5 항에 있어서,

상기 합성된 타겟 이미지의 특징값을 추출한 후 분류 모델에 의해 타겟 레이블로 분류되도록 학습하는 타겟 속성 선택부를 더 포함하는 장치.

발명의 설명

기 술 분 야

[0001] 본 발명은 분류 결과 설명이 가능한 반 사실적 맵 생성 방법 및 그 장치에 관한 것이다.

배 경 기 술

[0003] 설명 가능한 인공지능이란 인공지능이 내린 결정이나 답을 인공지능 스스로 사람이 이해할 수 있는 수준의 형태로 설명하는 방법으로, 의사결정에 대한 인과관계를 분석하여 적절한 근거를 사용자에게 제공해줌으로써 투명성과 신뢰성을 확보할 수 있도록 할 수 있다.

[0004] 특히, 딥 러닝이 다양한 분야에서 월등한 성능을 도출해내면서 이에 대응되는 해석 가능성(Interpretability)과 설명 가능성(Explainability)에 대한 요구와 함께 연구가 활발히 이루어지고 있다.

[0005] 하지만 의사결정에 대한 해석성과 높은 성능을 동시에 달성하는 것은 명백한 음의 상관관계로 인해 설명 가능한 인공지능 분야에서도 여전히 해결되지 않은 문제이다.

[0006] 사전 학습된 모델로부터 역전파(Back propagation)를 활용한 기울기 기반 방법이나 입력에 휴리스틱(Heuristic) 또는 섭동을 적용하여 모델을 해석하는 다양한 연구들이 진행되었지만 예측 결과를 해석하기 위한 근본적인 설명성을 유도하기에는 한계가 존재한다.

발명의 내용

해결하려는 과제

[0008] 본 발명은 분류 결과 설명이 가능한 반 사실적 맵 생성 방법 및 그 장치를 제공하기 위한 것이다.

[0009] 또한, 본 발명은 입력을 타겟 클래스로 분류하기 위한 맵을 시각화하여 인과 관계 분석을 위한 근거로 제공할 수 있는 분류 결과 설명이 가능한 반 사실적 맵 생성 방법 및 그 장치를 제공하기 위한 것이다.

[0010] 또한, 본 발명은 입력 샘플의 구조적 특징을 유지하며 타겟 클래스에 의해 부여된 속성이 조건화된 레이블에 기반하여 타겟 클래스로 분류하기 위한 맵을 생성하여 시각화할 수 있는 분류 결과 설명이 가능한 반 사실적 맵 생성 방법 및 그 장치를 제공하기 위한 것이다.

과제의 해결 수단

[0012] 본 발명의 일 측면에 따르면, 분류 결과 설명이 가능한 반 사실적 맵 생성 방법이 제공된다.

[0013] 본 발명의 일 실시예에 따르면, 입력 이미지와 타겟 레이블을 신경망 기반 반 사실적 맵 생성 모델에 적용하여 반 사실적 맵(counterfactual map)을 생성하는 단계; 상기 입력 이미지와 상기 반 사실적 맵간의 선형 합을 통해 합성된 타겟 이미지를 생성하는 단계; 및 상기 합성된 타겟 이미지의 특징값을 추출한 후 분류 모델에 의해

타겟 레이블로 분류되도록 학습하는 단계를 포함하는 분류 결과 설명이 가능한 반 사실적 맵 생성 방법이 제공될 수 있다.

[0014] 상기 반 사실적 맵 생성 모델은, 상기 입력 이미지의 구조적 정보를 유지하도록 주기 일관성 손실을 고려하여 상기 입력 이미지에 의존적인 반 사실적 맵을 생성할 수 있다.

[0015] 상기 반 사실적 맵 생성 모델은 적대적 생성 신경망 기반 모델이되, 상기 합성된 타겟 이미지가 실제 샘플에 매핑되도록 판별부를 적대적 학습시키고, 상기 합성된 타겟 이미지가 타겟 레이블로 예측되기 위한 반 사실적 맵을 생성하도록 학습될 수 있다.

[0016] 상기 타겟 레이블은 입력 이미지를 타겟으로 변환하기 위한 속성이 부여된 레이블이다.

[0018] 본 발명의 다른 측면에 따르면, 분류 결과 설명이 가능한 반 사실적 맵 생성이 가능한 장치가 제공된다.

[0019] 본 발명의 일 실시예에 따르면, 입력 이미지를 분석하여 특징값을 추출하는 특징 추출부; 상기 입력 이미지의 특징값들과 타겟 레이블을 신경망 기반 반 사실적 맵 생성 모델에 적용하여 반 사실적 맵(counterfactual map)을 생성하는 맵 생성부; 상기 입력 이미지와 상기 반 사실적 맵간의 선형 합을 통해 합성된 타겟 이미지를 생성하는 타겟 이미지 생성부를 포함하는 장치가 제공될 수 있다.

[0020] 상기 반 사실적 맵 생성 모델은 적대적 생성 신경망 기반 모델이되, 상기 합성된 타겟 이미지가 실제 샘플에 매핑되도록 판별부를 적대적 학습하며, 상기 합성된 타겟 이미지가 타겟 레이블로 예측되기 위한 반 사실적 맵을 생성하도록 학습될 수 있다.

[0021] 상기 합성된 타겟 이미지의 특징값을 추출한 후 분류 모델에 의해 타겟 레이블로 분류되도록 학습하는 타겟 속성 선택부를 더 포함할 수 있다.

발명의 효과

[0023] 본 발명의 일 실시예에 따른 분류 결과 설명이 가능한 반 사실적 맵 생성 방법 및 그 장치를 제공함으로써 입력을 타겟 클래스로 분류하기 위한 맵을 시각화하여 인과 관계 분석을 위한 근거로 제공할 수 있다.

[0024] 또한, 본 발명은 입력 샘플의 구조적 특징을 유지하며 타겟 클래스에 의해 부여된 속성이 조건화된 레이블에 기반하여 타겟 클래스로 분류하기 위한 맵을 생성하여 시각화할 수 있는 이점이 있다.

도면의 간단한 설명

[0026] 도 1은 본 발명의 일 실시예에 따른 분류 결과 설명이 가능한 맵 생성 장치의 내부 구성을 개략적으로 도시한 도면.

도 2는 본 발명의 일 실시예에 따른 반 사실적 맵 생성을 설명하기 위해 도시한 도면.

도 3은 본 발명의 일 실시예에 따른 반 사실적 맵 생성을 위한 구조를 도시한 도면.

도 4는 본 발명의 일 실시예에 따른 반 사실적 맵 생성 모델의 상세 구조를 도시한 도면.

도 5는 본 발명의 일 실시예에 따른 입력 이미지와 반 사실적 맵간의 합성 결과를 예시한 도면.

도 6은 본 발명의 일 실시예에 따른 분류 결과 설명이 가능한 반 사실적 맵 생성 방법을 나타낸 순서도.

발명을 실시하기 위한 구체적인 내용

[0027] 본 명세서에서 사용되는 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "구성된다" 또는 "포함한다" 등의 용어는 명세서상에 기재된 여러 구성 요소들, 또는 여러 단계들을 반드시 모두 포함하는 것으로 해석되지 않아야 하며, 그 중 일부 구성 요소들 또는 일부 단계들은 포함되지 않을 수도 있고, 또는 추가적인 구성 요소 또는 단계들을 더 포함할 수 있는 것으로 해석되어야 한다. 또한, 명세서에 기재된 "...부", "모듈" 등의 용어는 적어도 하나의 기능이나 동작을 처리하는 단위를 의미하며, 이는

하드웨어 또는 소프트웨어로 구현되거나 하드웨어와 소프트웨어의 결합으로 구현될 수 있다.

[0028] 이하, 첨부된 도면들을 참조하여 본 발명의 실시예를 상세히 설명한다.

[0030] 도 1은 본 발명의 일 실시예에 따른 분류 결과 설명이 가능한 맵 생성 장치의 내부 구성을 개략적으로 도시한 도면이고, 도 2는 본 발명의 일 실시예에 따른 반 사실적 맵 생성을 설명하기 위해 도시한 도면이며, 도 3은 본 발명의 일 실시예에 따른 반 사실적 맵 생성을 위한 구조를 도시한 도면이고, 도 4는 본 발명의 일 실시예에 따른 반 사실적 맵 생성 모델의 상세 구조를 도시한 도면이고, 도 5는 본 발명의 일 실시예에 따른 이미지와 반 사실적 맵간의 합성 결과를 예시한 도면이다.

[0031] 도 1을 참조하면, 본 발명의 일 실시예에 따른 장치(100)는 반 사실적 맵 생성 모델(110), 타겟 속성 선택부(120), 메모리(130) 및 프로세서(140)를 포함하여 구성된다.

[0032] 반 사실적 맵 생성 모델(110)은 입력 이미지와 타겟 레이블을 입력받아 입력 이미지의 구조적 정보를 고려한 반 사실적 맵(counterfactual map)을 생성하기 위한 수단이다.

[0033] 예를 들어, 입력 이미지가 정상적인 뇌 이미지라고 가정하며, 타겟 레이블을 치매 레이블이라고 가정하기로 한다. 반 사실적 맵 생성 모델(110)은 정상적인 뇌 이미지의 구조적 특징을 고려하여 치매 레이블이 반영된 반 사실적 맵을 생성할 수 있다.

[0034] 이해와 설명의 편의를 도모하기 위해, 도 2를 가정하기로 한다.

[0035] 도 2의 210과 같은 입력 이미지(예를 들어, 숫자 1에 대한 이미지)가 입력되고, 타겟 레이블이 숫자 "0"에 대한 레이블이라고 가정하기로 한다.

[0036] 반 사실적 맵 생성 모델(110)은 도 2의 220과 같이 입력 이미지의 구조적 특징이 고려된 타겟 레이블이 반영된 반 사실적 맵을 생성할 수 있다.

[0037] 이에 대해 보다 상세히 설명하기로 한다.

[0038] 본 발명의 일 실시예에 따르면, 반 사실적 맵 생성 모델(110)은 도 4에 도시된 바와 같이, 인코더(410), 생성부(420) 및 판별부(430)를 포함하여 구성된다. 또한, 반 사실적 맵 생성 모델(110)은 클래스 예측기(440)을 포함할 수 있다.

[0039] 도 3에는 반 사실적 맵 생성 모델(110)의 일부 구조가 예시되어 있다. 반 사실적 맵 생성 모델의 인코더($\mathcal{E}_\theta$)와 생성부($\mathcal{G}_\phi$)는 도 3에 도시된 바와 같이, 스킵 연결(skip connections)로 통합된(concatenated) 타일화된 타겟 레이블을 가지는 U-NET 변형으로 설계된다. 생성부($\mathcal{G}_\phi$)는 다방향 반 사실 추론이 가능하도록 반 사실적 맵을 합성할 수 있다.

[0040] 우선 이해와 설명의 편의를 도모하기 위해, $P = \{P_{\mathbf{x}}, P_{\mathbf{y}}\}$를 데이터 셋으로 정의하기로 한다. 본 명세서에서 $\{\mathbf{x}, \mathbf{y}\} \sim P_{\mathbf{x},\mathbf{y}}$는 페어(pair)를 이루는 데이터와 레이블인 것으로 정의하기로 하며, $\mathbf{x} \sim P_{\mathbf{x}}, \mathbf{y} \sim P_{\mathbf{y}}$는 페어를 이루지 않은 데이터와 레이블인 것으로 정의하기로 한다.

[0041] 반 사실적 맵은 분류 모델($\mathcal{F}$)의 반 사실적 설명을 생성할 수 있는 관심 영역 맵(salient map)이다. $\mathbf{M}_{\mathbf{x}}$는 다 방향 반 사실적 맵으로 정의하기로 한다. $\mathbf{x}$를 입력 샘플이라 가정하기로 한다. 반 사실적 맵($\mathbf{M}_{\mathbf{x}}$)은 입력 샘플($\mathbf{x}$)이 $\tilde{\mathbf{x}} = \mathbf{x} + \mathbf{M}_{\mathbf{x}}$와 같이 추가되는 경우, 분류 모델($\mathcal{F}$)은 원래 레이블에 대한 다른 레이블 보완으로 분류하도록 하는 맵이며, 그 역도 동일하다.

[0042] 따라서, 속성을 가지는 반 사실적 맵을 조건화할 수 있다면, 입력 샘플(입력 이미지)를 관심 있는 다른 타겟으

로 변경할 수 있다.

[0043] 이를 수학식으로 나타내면, 수학식 1과 같다.

수학식 1

[0044] $$\tilde{\mathbf{x}} := \mathbf{x} + \mathbf{M}_{\mathbf{x},\mathbf{y}} \quad \mathcal{F}(\tilde{\mathbf{x}}) = \mathbf{y}$$

[0045] 여기서, x는 입력 데이터를 나타내고, $\mathbf{y}$ 는 타겟 레이블을 나타내고, x와 y는 쌍을 이루고 있지 않는다. 즉, $\mathbf{x} \sim P_{\mathbf{x}}, \mathbf{y} \sim P_{\mathbf{y}}$ 이다.

[0046] 반 사실적 맵 생성 모델(110)은 입력 이미지와 타겟 클래스에 상응하는 속성이 부여된 타겟 레이블을 입력받은 후 입력 이미지의 구조적 정보를 유지하는 타겟 레이블이 반영된 반 사실적 맵을 생성할 수 있다.

[0047] 반 사실적 맵 생성 모델(110)은 입력 이미지에 대해 타겟 레이블($\mathbf{y}$)로 조건화된 반 사실적 맵을 통합할 수 있는 조건부 적대적 생성 신경망(GAN)의 변형이다.

[0048] 반 사실적 맵은 수학식 2와 같이 나타낼 수 있다.

수학식 2

[0049] $$\mathbf{M}_{\mathbf{x},\mathbf{y}} = \mathcal{G}_{\phi}\left(\mathcal{E}_{\theta}(\mathbf{x}), \mathbf{y}\right)$$

[0050] 즉, 반 사실적 맵은 입력 이미지에서 추출된 특징값들과 타겟 레이블을 이용하여 입력 이미지의 구조적 정보를 보존하도록 생성될 수 있다.

[0051] 적대적 생성 신경망(GAN)의 일반적인 학습 스킴에 따라 판별부($\mathcal{D}_{\psi}$)는 합성된 타겟 이미지($\tilde{\mathbf{x}}$)로부터 실제 이미지들($\hat{\mathbf{x}} \sim P_{\mathbf{x}}$)을 구별하도록 학습될 수 있다.

[0052] 반 사실적 추론은 가상과 주어진 현실 사이에 적절한 균형을 필요로 한다.

[0053] 이하에서는 반 사실적 맵 생성 모델의 손실 함수에 대해 각각 설명하기로 한다.

[0054] (1) 적대적 손실 함수

[0055] 반 사실적 맵 생성 모델은 적대적 훈련 중 안정성으로 인해 최소 제곱 GAN(LSGAN) 목적 함수를 기반으로 적대적 손실 함수를 정의할 수 있다.

[0056] LSGAN의 목적은 판별부의 결정 경계에서 멀리 떨어진 샘플에 패널티를 적용하여 안정적인 반 사실적 맵 생성 모델의 학습이 가능케 하는 것이다. 적대적 손실 함수는 생성된 반 사실적 맵이 입력 샘플들을 파괴하거나 타겟 속성을 무시하지 않도록 할 수 있다. 즉, 실제 이미지(샘플)과 가짜 샘플 사이에 적절한 균형을 포함하도록 할 수 있다.

[0057] 분류부와 생성부의 적대적 손실 함수는 수학식 3 및 수학식 4와 같이 나타낼 수 있다.

수학식 3

[0058]
$$\mathcal{L}_{adv}^{\mathcal{D}_\psi} = \frac{1}{2}\mathbb{E}_{\hat{\mathbf{x}}\sim P_{\mathbf{x}}}\left[(D_\psi(\hat{\mathbf{x}})-1)^2\right] + \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim P_{\mathbf{x}},\mathbf{y}\sim P_{\mathbf{y}}}\left[(D_\psi(\tilde{\mathbf{x}}))^2\right]$$

수학식 4

[0060]
$$\mathcal{L}_{adv}^{\mathcal{G}_\phi} = \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim P_{\mathbf{x}},\mathbf{y}\sim P_{\mathbf{y}}}\left[(D_\psi(\tilde{\mathbf{x}})-1)^2\right]$$

[0062] (2) 주기 일관성 손실

[0063] 주기 일관성 손실은 더 나은 다-"뉨* 반 사실적 맵을 생성하기 위해 이용된다. 단방향 반 사실적 맵 생성 시나리오에서 실제 샘플 $\hat{\mathbf{x}}$ 과 합성된 샘플$\tilde{\mathbf{x}}$은 항상 동일한 특정 분류(class)에서 샘플링된다. 실제 및 가짜 샘플에는 항상 지정된 레이블이 있으므로, 이 설정에는 주기 일관성 손실이 필요치 않다.

[0064] 그러나 판별부는 실제 또는 가짜 샘플만 분류하기 때문에 생성부를 안내하여 다 방향 반 사실적 맵을 생성할 수 있는 능력이 없다. 따라서, 본 발명의 일 실시예에서는 순방향 사이클이 예를 들어, $\mathbf{M}_{\mathbf{x},\mathbf{y}}$는 쌍을 이루지 않는 데이터 $\mathbf{x} \sim P_{\mathbf{x}}$ 와 레이블 $\mathbf{y} \sim P_{\mathbf{y}}$ 로부터 생성되는 것과 같은 임의의 조건을 갖는 반 사실적 맵을 생성하도록 주기 일관성 손실을 고려하기로 한다.

[0065] 역방향 주기는 입력 샘플의 레이블, 쌍을 이루는 데이터와 레이블($(\mathbf{x},\mathbf{y}') \sim P_{\mathbf{x},\mathbf{y}}$)로부터 생성된 $\mathbf{M}_{\tilde{\mathbf{x}},\mathbf{y}'}$ 에 조건이 지정된 맵을 생성할 수 있다.

[0066] 주기 일관성 손실 함수는 수학식 5와 같이 나타낼 수 있다.

수학식 5

[0067]
$$\mathcal{L}_{cyc} = \mathbb{E}_{(\mathbf{x},\mathbf{y}')\sim P_{\mathbf{x},\mathbf{y}},\,\mathbf{y}\sim P_{\mathbf{y}}}\left[\left\|(\tilde{\mathbf{x}}+\mathbf{M}_{\tilde{\mathbf{x}},\mathbf{y}'})-\mathbf{x}\right\|_1\right],$$

[0068] 여기서, $\mathbf{M}_{\tilde{\mathbf{x}},\mathbf{y}'} = \mathcal{G}_\phi\left(\mathcal{E}_\theta(\tilde{\mathbf{x}}), \mathbf{y}'\right)$ 이다.

[0070] (3) 클래스 예측 손실

[0071] 타겟 속성 선택부(120)는 합성된 반 사실적 맵에 대한 속성 선택을 시행하는데 반 사실적 맵 생성 모델(110)과 보완적인 방식으로 동작한다.

[0072] 타겟 속성 선택부(120)의 목표는 반 사실적 맵 생성 모델(110)의 입력 샘플을 타겟 레이블로 분류하도록 변환하는 반 사실적 맵을 생성하도록 가이드하는 것이다.

수학식 6

[0073] $$\mathcal{L}_{cls} = \mathbb{E}_{\mathbf{x}\sim P_{\mathbf{x}},\mathbf{y}\sim P_{\mathbf{y}}}\left[CE(\mathbf{y}, \tilde{\mathbf{y}})\right]$$

[0074] 여기서, y는 타겟 레이블을 나타내고, $\tilde{\mathbf{y}} = \mathcal{F}(\tilde{\mathbf{x}})$, CE 는 크로스-엔트로피 함수를 나타낸다.

[0075] 개념적으로 타겟 속성 선택부(120)의 역할은 GAN의 판별부의 역할과 유사하나 그 목적이 상이하다. 판별부는 실제 샘플과 가짜 샘플을 구별하는 방법을 학습하는 반면, 타겟 속성 선택부는 입력 샘플을 분류하도록 사전 학습된다. 판별부는 보다 현실적인 샘플을 생성하기 위해 생성기와 함께 minmax 게임을 수행하는 반면, 타겟 속성 선택부는 반 사실적 맵 생성 모델(110)이 클래스별 샘플을 생성할 수 있도록 결정론적 지침을 제공한다.

[0077] (4) 반 사실적 맵 손실

[0078] 반 사실적 맵 손실은 증가하는 반 사실적 맵의 값을 제한한다.

수학식 7

[0079] $$\mathcal{L}_{map} = \mathbb{E}_{\mathbf{x}\sim P_{\mathbf{x}},\mathbf{y}\sim P_{\mathbf{y}}}\left[\lambda_1 \|\mathbf{M}_{\mathbf{x},\mathbf{y}}\|_1 + \lambda_2 \|\mathbf{M}_{\mathbf{x},\mathbf{y}}\|_2\right]$$

[0080] 여기서, λ_1 와 λ_2 는 초매개변수로서 가중치 상수값을 나타낸다.

[0081] 반 사실적 맵 손실은 반 사실적 맵 생성에서 두가지 문제를 해결하는데 중요하다. 생성 모델의 경우, 생성자가 단순히 판별자를 속이기 위한 특정 픽셀 하나 또는 작은 영역의 섭동을 통해 모델의 결정을 변경할 수 있는데 이는 적대적 공격과 관련이 있다. 즉, 입력 이미지의 아이덴티티를 전혀 고려하지 않은 맵이 생성되기 때문에 맵 손실이 필요하다.

[0082] 그리고, 최종 사용자의 관점에서 가장 중요한 특징들만 반 사실적 설명으로 나타내도록 표현되어야 한다. 따라서, 반 사실적 맵의 크기에 제약을 두면 상술한 두가지 문제를 해결할 수 있다.

[0083] 상술한 바와 같은 적대적 손실 함수, 주기 일관성 손실 함수, 타겟 속성 선택 손실 함수와 반 사실적 맵 손실 함수를 모두 고려하여 손실 함수가 수학식 8과 같이 정의될 수 있다.

수학식 8

[0084] $$\mathcal{L} = \lambda_3 \mathcal{L}_{adv}^{\mathcal{D}_\psi} + \lambda_4 \mathcal{L}_{adv}^{\mathcal{G}_\phi} + \lambda_5 \mathcal{L}_{cyc} + \lambda_6 \mathcal{L}_{cls} + \mathcal{L}_{map}$$

[0085] 여기서, λ는 초매개변수를 나타낸다.

[0086] 훈련 동안, 생성 프로세스 전체에서 속성이 일관되게 유지되도록 반 사실적 맵 생성 모델(110)의 인코더의 고정된 가중치를 타겟 속성 선택부(120)와 공유할 수 있다.

[0087] 타겟 속성 선택부(120)는 타겟 이미지의 특징값을 추출한 후 분류 모델을 통해 레이블로 분류하기 위한 수단이다.

[0088] 예를 들어, 타겟 속성 선택부(120)는 타겟 이미지의 특징을 추출하기 위해 반 사실적 맵 생성 모델(110)에서 사용하는 특징 추출기를 동일하게 사용하며, 분류 모델을 통해 타겟 레이블로 분류되도록 학습된다.

[0089] 도 5에는 입력 이미지의 구조적 정보를 유지하여 타겟 레이블이 반영된 반 사실적 맵이 생성된 일 예들이 예시되어 있으며, 이를 다시 입력 이미지와의 합성을 통해 타겟 이미지를 생성한 결과가 예시되어 있다.

[0090] 메모리(130)는 본 발명의 일 실시예에 따른 분류 결과 설명이 가능한 반 사실적 맵 생성 방법을 위한 다양한 명령어들 저장하기 위한 수단이다.

[0091] 프로세서(140)는 본 발명의 일 실시예에 따른 장치(100)의 내부 구성 요소들(예를 들어, 반 사실적 맵 생성 모델(110), 타겟 속성 선택부(120), 메모리(130) 등)을 제어하기 위한 수단이다.

[0092]

[0093] 도 6은 본 발명의 일 실시예에 따른 분류 결과 설명이 가능한 반 사실적 맵 생성 방법을 나타낸 순서도이다.

[0094] 단계 610에서 장치(100)는 입력 이미지를 분석하여 특징값들을 추출한다.

[0095] 단계 615에서 장치(100)는 타겟 클래스의 속성이 부여된 타겟 레이블을 입력받는다.

[0096] 단계 620에서 장치(100)는 입력 이미지의 특징값들과 타겟 레이블을 반 사실적 맵 생성 모델에 적용하여 입력 이미지의 구조적 정보를 유지하면서 타겟 레이블의 속성이 반영된 반 사실적 맵을 생성한다.

[0097] 단계 625에서 장치(100)는 입력 이미지와 반 사실적 맵간의 선형합을 통한 합성된 타겟 이미지를 생성한다.

[0098] 단계 630에서 장치(100)는 합성된 타겟 이미지가 실제 샘플들에 매핑되도록 적대적 학습하며, 타겟 이미지가 부여된 타겟 레이블로 예측되기 위한 특징들을 반영한 반 사실적 맵을 정확하게 생성하도록 학습된다.

[0099] 단계 635에서 장치(100)는 합성된 타겟 이미지의 특징값을 추출한 후 분류 모델에 의해 타겟 레이블로 분류되도록 학습한다.

[0100] 전술한 바와 같이, 본 발명의 일 실시예에 따르면 입력 샘플(이미지)의 구조적 정보가 반영된 타겟 레이블에 의존적인 특징들이 반영된 반 사실적 맵을 생성할 수 있다. 이와 같이 시각화된 반 사실적 맵을 생성하여 제공함으로써 최종 사용자가 분류 모델의 분류 결과의 인과관계를 분석할 수 있도록 근거를 제공하며 인간 지식과 유사한 수준의 설명성을 도출하도록 할 수 있는 이점이 있다.

[0101] 본 발명의 실시 예에 따른 장치 및 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 컴퓨터 판독 가능 매체에 기록되는 프로그램 명령은 본 발명을 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 분야 통상의 기술자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media) 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴

파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.

[0102] 상술한 하드웨어 장치는 본 발명의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0104] 이제까지 본 발명에 대하여 그 실시 예들을 중심으로 살펴보았다. 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자는 본 발명이 본 발명의 본질적인 특성에서 벗어나지 않는 범위에서 변형된 형태로 구현될 수 있음을 이해할 수 있을 것이다. 그러므로 개시된 실시 예들은 한정적인 관점이 아니라 설명적인 관점에서 고려되어야 한다. 본 발명의 범위는 전술한 설명이 아니라 특허청구범위에 나타나 있으며, 그와 동등한 범위 내에 있는 모든 차이점은 본 발명에 포함된 것으로 해석되어야 할 것이다.

부호의 설명

[0106] 100: 장치

110: 반 사실적 맵 생성 모델

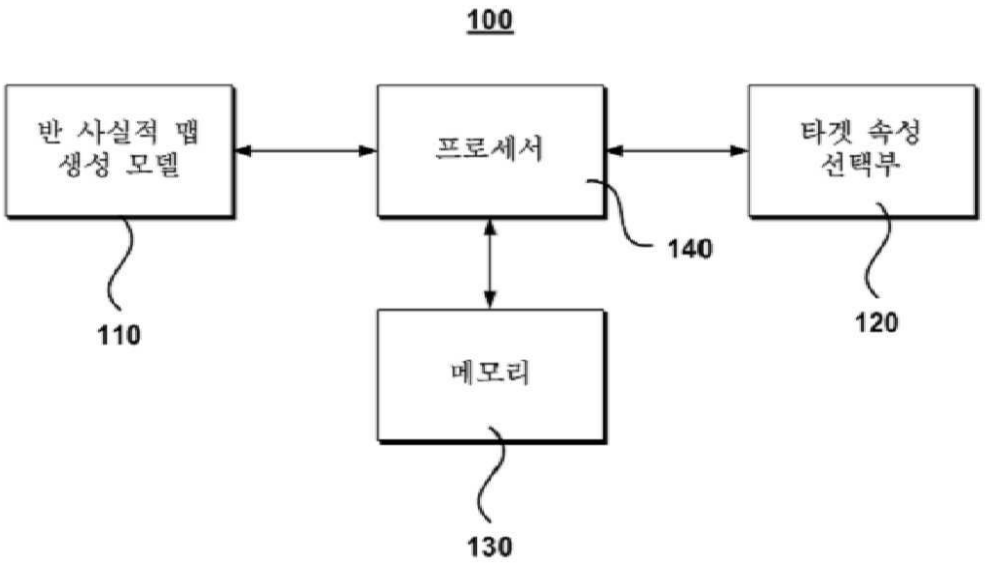
120: 타겟 속성 선택부

130: 메모리

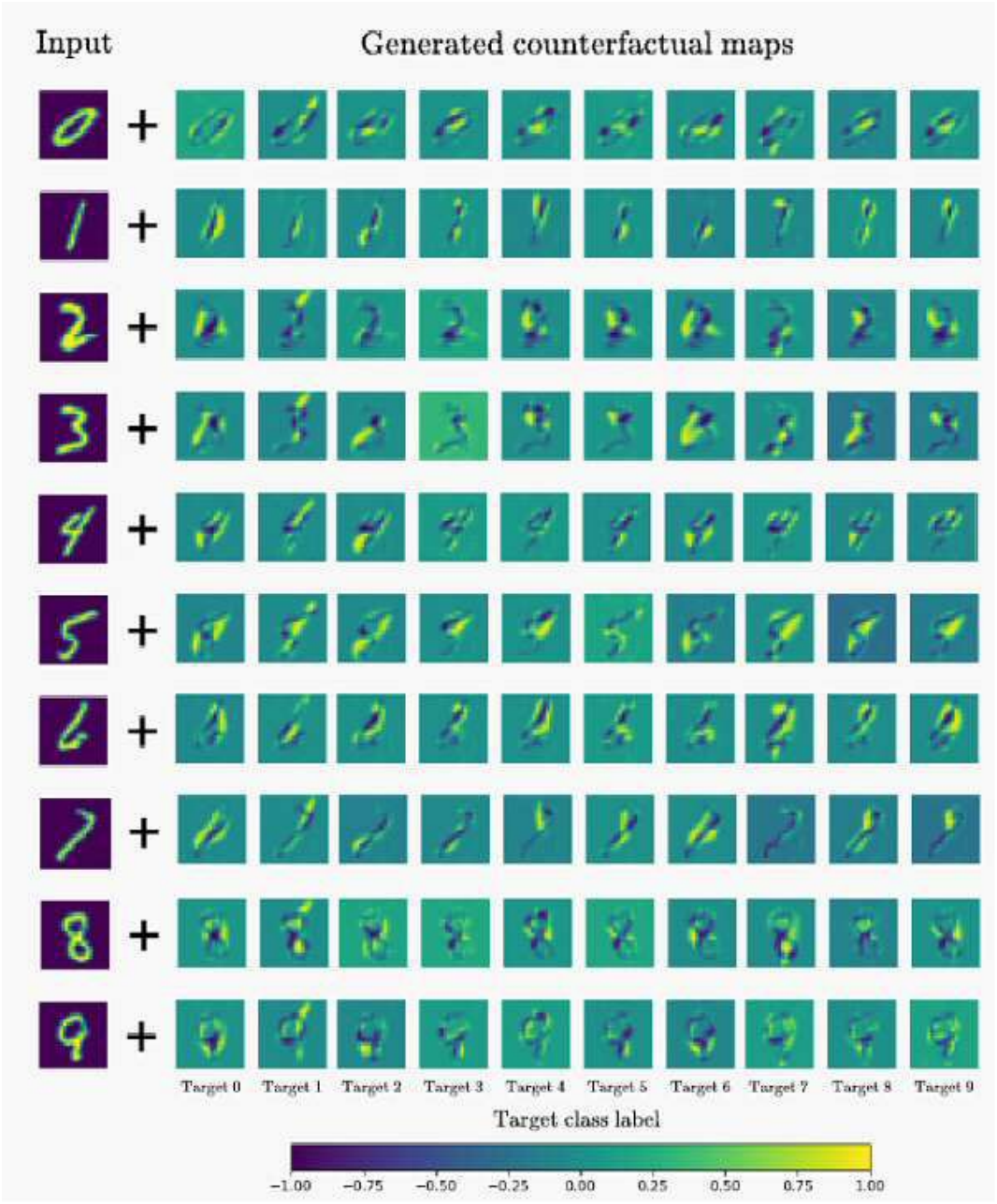
140: 프로세서

도면

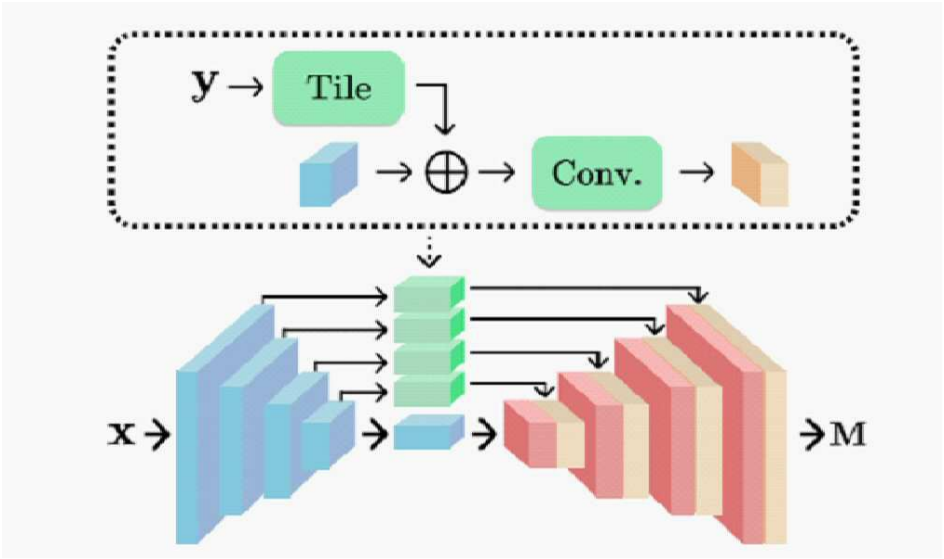
도면1



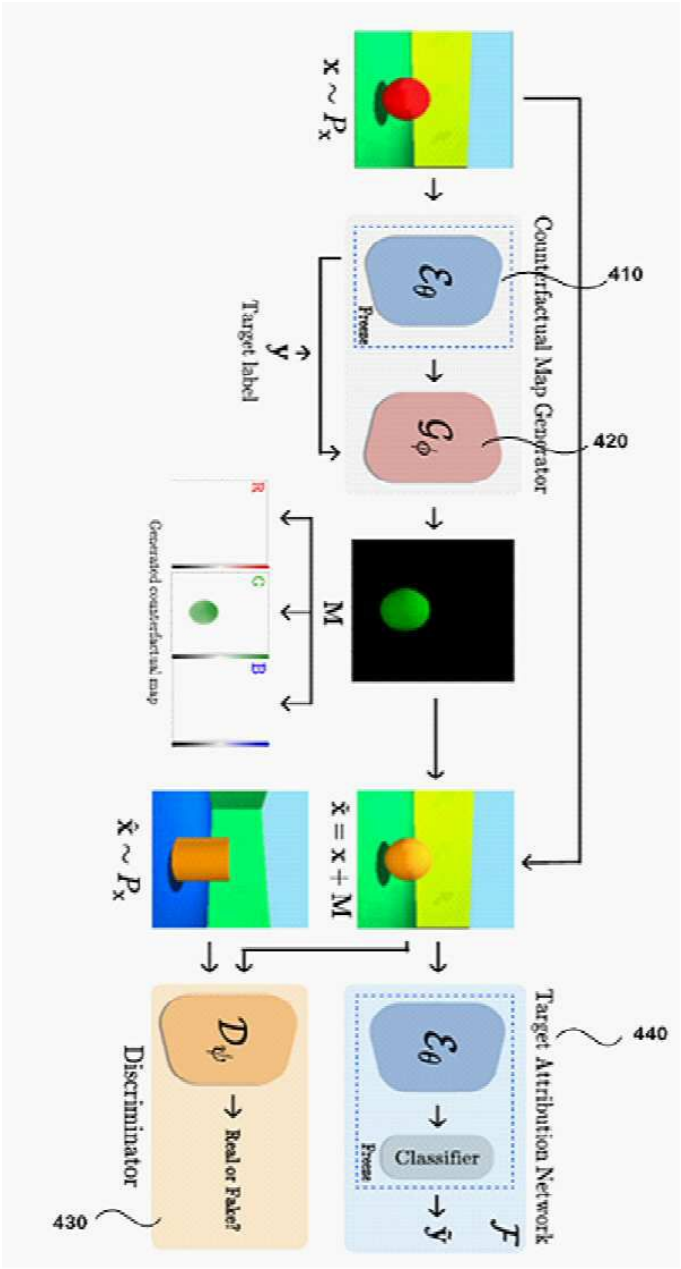
도면2


Input
Generated counterfactual maps
Target 0
Target 1
Target 2
Target 3
Target 4
Target 5
Target 6
Target 7
Target 8
Target 9
Target class label
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00

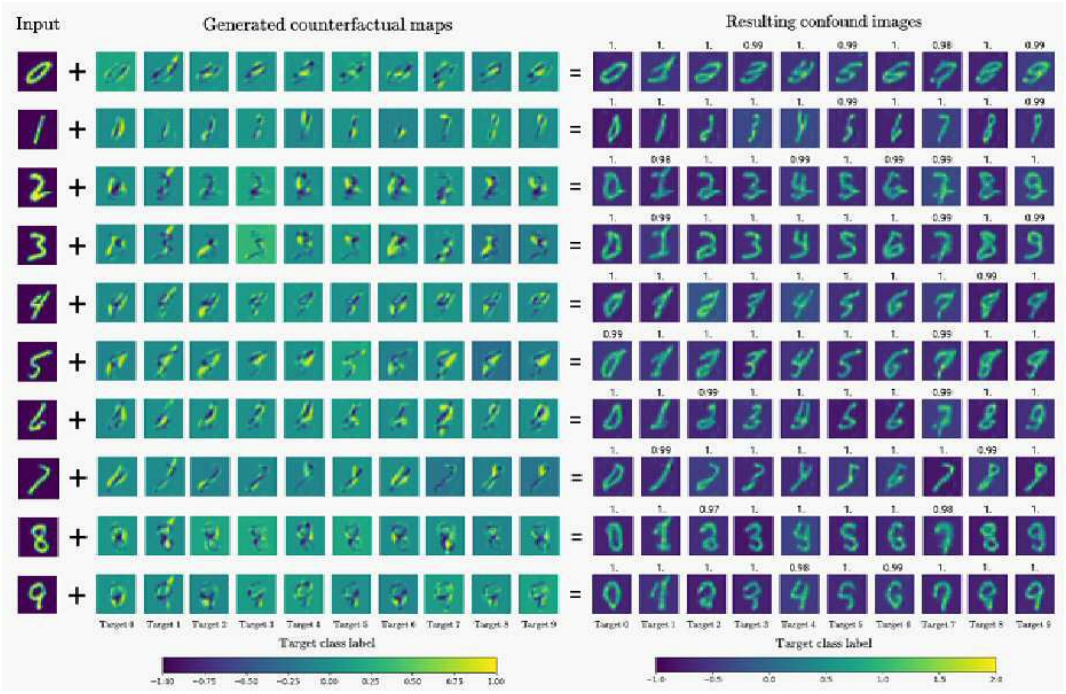
도면3



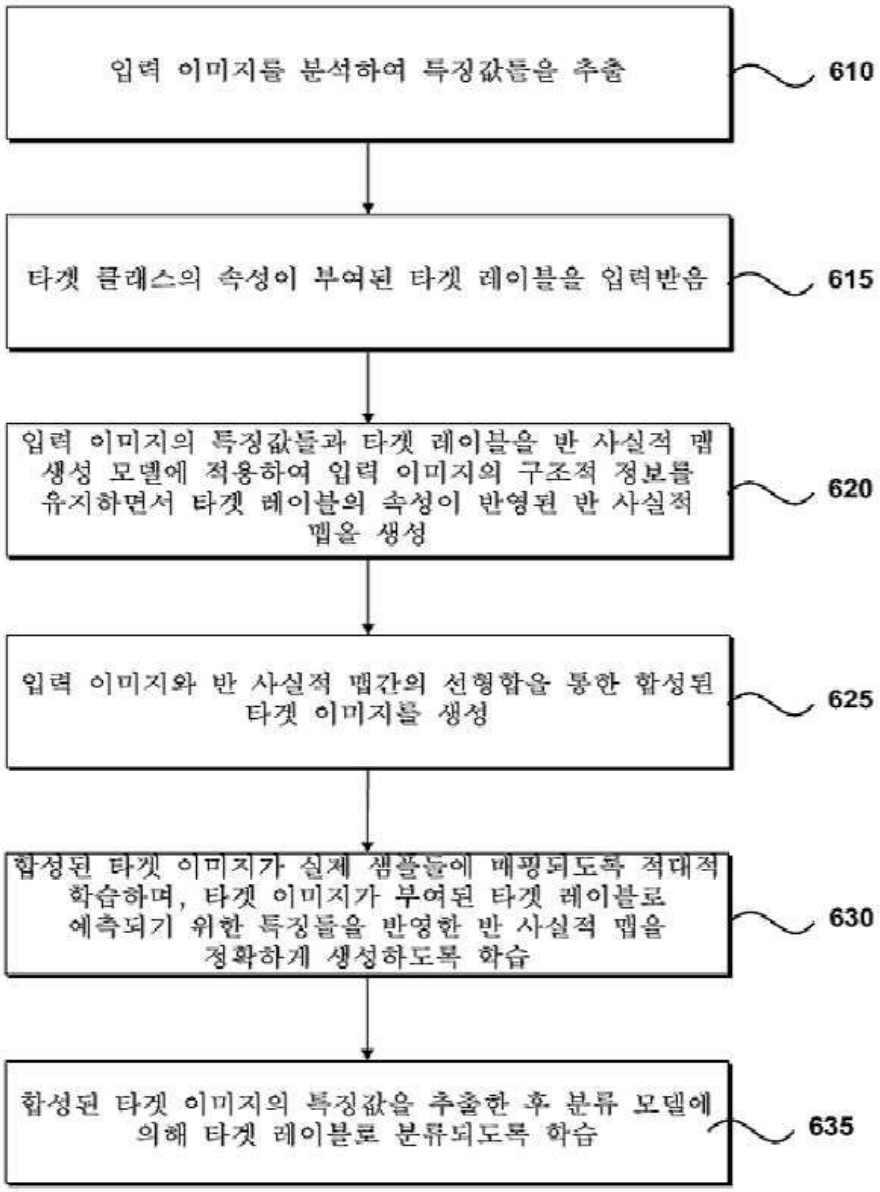
도면4


$x \sim P_x$
$\mathcal{E}_\theta$
Freeze
Counterfactual Map Generator
410
Target label
y
$\mathcal{G}_\phi$
420
R
G
B
Generated counterfactual map
M
$\tilde{x} = x + M$
$\tilde{x} \sim P_x$
$\mathcal{E}_\theta$
Classifier
Freeze
$\mathcal{F}$
$\tilde{y}$
Target Attribution Network
440
$\mathcal{D}_\psi$
Real or Fake?
Discriminator
430

도면5


Input
Generated counterfactual maps
Resulting confound images
Target class label
Target class label

도면6


입력 이미지를 분석하여 특징값들을 추출
610
타겟 클래스의 속성이 부여된 타겟 레이블을 입력받음
615
입력 이미지의 특징값들과 타겟 레이블을 반 사실적 맵 생성 모델에 적용하여 입력 이미지의 구조적 정보를 유지하면서 타겟 레이블의 속성이 반영된 반 사실적 맵을 생성
620
입력 이미지와 반 사실적 맵간의 선형합을 통한 합성된 타겟 이미지를 생성
625
합성된 타겟 이미지가 실제 샘플들에 매핑되도록 적대적 학습하며, 타겟 이미지가 부여된 타겟 레이블로 예측되기 위한 특징들을 반영한 반 사실적 맵을 정확하게 생성하도록 학습
630
합성된 타겟 이미지의 특징값을 추출한 후 분류 모델에 의해 타겟 레이블로 분류되도록 학습
635