



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0029346
(43) 공개일자 2025년03월05일

(51) 국제특허분류(Int. Cl.)

G06N 5/04 (2023.01) G06N 3/0455 (2023.01)
G06N 3/0475 (2023.01) G06N 3/08 (2023.01)
G06N 3/094 (2023.01)

(52) CPC특허분류

G06N 5/045 (2023.01)
G06N 3/0455 (2023.01)

(21) 출원번호 10-2023-0109971

(22) 출원일자 2023년08월22일

심사청구일자 2023년08월22일

(71) 출원인

고려대학교 산학협력단

서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)

(72) 발명자

이성환

서울특별시 강남구 언주로30길 13, 비동 2902호
(도곡동, 대림아크로빌, 대림아크로텔)

나승협

서울특별시 동작구 여의대방로 28 (신대방동 현대아파트) 102동 1902호

(74) 대리인

윤귀상

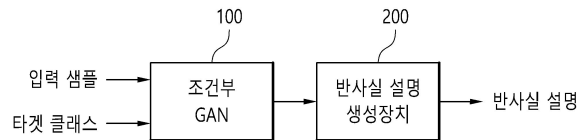
전체 청구항 수 : 총 13 항

(54) 발명의 명칭 분류기의 판단에 대한 반사실 설명 생성 방법, 장치 및 기록 매체

(57) 요약

본 발명은 분류기의 판단에 대한 반사실 설명 생성 방법에 대한 것으로, 입력 샘플과 분류기(classifier)에 대한 타겟 클래스 정보에 대한 참조 반사실 정보를 수신하고, 상기 참조 반사실 정보로부터 상기 입력 샘플과 참조 반사실 정보에 대한 잠재 벡터를 도출하고, 상기 입력 샘플의 잠재 벡터를 상기 참조 반사실 정보의 잠재 벡터에 대해 변환시키면서 상기 분류기의 클래스 변경에 대한 결정 경계(Decision Boundary)를 확인하고, 상기 변환된 잠재 벡터가 적용된 입력 샘플이 상기 분류기의 결정 경계를 넘어서는 시점의 잠재 벡터를 검출하고, 검출된 잠재 벡터를 기반으로 반사실 설명을 생성한다.

대표도 - 도1



(52) CPC특허분류

G06N 3/0475 (2023.01)

G06N 3/088 (2023.01)

G06N 3/094 (2023.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711193973
과제번호	2019-0-00079-005
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신방송혁신인재양성
연구과제명	인공지능대학원지원(고려대학교)
기 여 율	1/2
과제수행기관명	고려대학교산학협력단
연구기간	2023.01.01 ~ 2023.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1711193255
과제번호	2022-0-00984-002
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	사람중심인공지능핵심원천기술개발
연구과제명	플러그앤플레이 방식으로 설명가능성을 제공하는 인공지능 기술 개발 및 인공지능
시스템에 대한 설명 제공 검증	
기 여 율	1/2
과제수행기관명	한국과학기술원
연구기간	2023.01.01 ~ 2023.12.31

명세서

청구범위

청구항 1

입력 샘플과 분류기(classifier)에 대한 타겟 클래스 정보에 대한 참조 반사실 정보를 수신하는 단계;

상기 참조 반사실 정보로부터 상기 입력 샘플과 참조 반사실 정보에 대한 잠재 벡터를 도출하는 단계;

상기 입력 샘플의 잠재 벡터를 상기 참조 반사실 정보의 잠재 벡터에 대해 변환시키면서 상기 분류기의 클래스 변경에 대한 결정 경계(Decision Boundary)를 확인하는 단계; 및

상기 변환된 잠재 벡터가 적용된 입력 샘플이 상기 분류기의 결정 경계를 넘어서는 시점의 잠재 벡터를 검출하고, 검출된 잠재 벡터를 기반으로 반사실 설명을 생성하는 단계;를 포함하는 분류기의 판단에 대한 반사실 설명 생성 방법.

청구항 2

제1항에 있어서,

상기 참조 반사실 정보는,

미리 마련된 조건부(conditional) 적대적 생성 신경망(GAN: Generative Adversarial Network)에 상기 입력 샘플과 타겟 클래스 정보를 입력하여 획득된 결과인, 분류기의 판단에 대한 반사실 설명 생성 방법.

청구항 3

제1항에 있어서,

상기 잠재 벡터를 도출하는 단계는,

상기 참조 반사실 정보에 적대적 생성 신경망(GAN: Generative Adversarial Network) 인버전(inversion)을 적용하여 상기 잠재 벡터를 도출하는 것인, 분류기의 판단에 대한 반사실 설명 생성 방법.

청구항 4

제1항에 있어서,

상기 분류기의 클래스 변경에 대한 결정 경계를 확인하는 단계는,

상기 입력 샘플에 대한 잠재 벡터를 상기 참조 반사실 정보에 대한 잠재 벡터 방향으로 변환시키면서 상기 결정 경계를 확인하되,

상기 입력 샘플에 대한 잠재 벡터를 상기 참조 반사실 정보에 대한 잠재 벡터 방향으로 변환시키는 것은, 상기 입력 샘플에 대한 잠재 벡터와 상기 참조 반사실 정보에 대한 잠재 벡터 사이에 선형 내삽을 진행하여 선형적으로 벡터 값을 변환하는 것임을 특징으로 하는, 분류기의 판단에 대한 반사실 설명 생성 방법.

청구항 5

제2항에 있어서,

상기 잠재 벡터 도출 단계 및 상기 결정 경계 확인 단계는 상기 조건부 GAN의 엔코더와 디코더 사이의 잠재 공간(Latent space)을 활용하여 수행되고, 상기 잠재 공간은 의미(semantic) 정보들이 얹히지 않고 풀린

(disentangled) 공간임을 특징으로 하는, 분류기의 판단에 대한 반사실 설명 생성 방법.

청구항 6

제1항에 있어서,

상기 반사실 설명은 상기 검출된 잠재 벡터를 디코딩하여 생성됨을 특징으로 하는, 분류기의 판단에 대한 반사실 설명 생성 방법.

청구항 7

입력 샘플과 분류기(classifier)에 대한 타겟 클래스 정보에 대한 참조 반사실 정보를 수신하는 수신부;

상기 참조 반사실 정보로부터 상기 입력 샘플과 참조 반사실 정보에 대한 잠재 벡터를 도출하는 잠재 벡터 도출부;

상기 입력 샘플의 잠재 벡터를 상기 참조 반사실 정보의 잠재 벡터에 대해 변환시키면서 상기 분류기의 클래스 변경에 대한 결정 경계(Decision Boundary)를 확인하는 의미정보 변환부;

상기 변환된 잠재 벡터가 적용된 입력 샘플이 상기 분류기의 결정 경계를 넘어서는 시점의 잠재 벡터를 검출하고, 검출된 잠재 벡터를 기반으로 반사실 설명을 생성하는 반사실 설명 생성부;를 포함하는 분류기의 판단에 대한 반사실 설명 생성 장치.

청구항 8

제7항에 있어서,

상기 참조 반사실 정보는,

미리 마련된 조건부(conditional) 적대적 생성 신경망(GAN: Generative Adversarial Network)에 상기 입력 샘플과 타겟 클래스 정보를 입력하여 획득된 결과인, 분류기의 판단에 대한 반사실 설명 생성 장치.

청구항 9

제7항에 있어서,

상기 잠재 벡터 도출부는,

상기 참조 반사실 정보에 적대적 생성 신경망(GAN: Generative Adversarial Network) 인버전(inversion)을 적용하여 상기 잠재 벡터를 도출하는 것인, 분류기의 판단에 대한 반사실 설명 생성 장치.

청구항 10

제7항에 있어서,

상기 의미정보 변환부는,

상기 입력 샘플에 대한 잠재 벡터를 상기 참조 반사실 정보에 대한 잠재 벡터 방향으로 변환시키면서 상기 결정 경계를 확인하되,

상기 입력 샘플에 대한 잠재 벡터를 상기 참조 반사실 정보에 대한 잠재 벡터 방향으로 변환시키는 것은, 상기 입력 샘플에 대한 잠재 벡터와 상기 참조 반사실 정보에 대한 잠재 벡터 사이에 선형 내삽을 진행하여 선형적으로 벡터 값을 변환하는 것임을 특징으로 하는, 분류기의 판단에 대한 반사실 설명 생성 장치.

청구항 11

제8항에 있어서,

상기 잠재 벡터 도출부 및 상기 의미정보 변환부는 상기 조건부 GAN의 엔코더와 디코더 사이의 잠재 공간(Latent space)을 활용하여 동작하며, 상기 잠재 공간은 의미(semantic) 정보들이 얹히지 않고 풀린(disentangled) 공간임을 특징으로 하는, 분류기의 판단에 대한 반사실 설명 생성 장치.

청구항 12

제7항에 있어서,

상기 반사실 설명 생성부는 상기 검출된 잠재 벡터를 디코딩하여 상기 반사실 설명을 생성함을 특징으로 하는, 분류기의 판단에 대한 반사실 설명 생성 장치.

청구항 13

청구항 제1항의 분류기의 판단에 대한 반사실 설명 생성 방법을 수행하기 위한 컴퓨터 프로그램이 기록된 기록 매체.

발명의 설명

기술 분야

[0001] 본 발명은 분류기의 판단에 대한 반사실 설명(Counterfactual explanation)을 생성하는 방법, 장치 및 기록 매체에 관한 것이다.

배경 기술

[0003] 근래에 인공지능 모델은 퍼셉트론(perceptron)에 기반한 신경망(Neural network) 구조로 큰 향상을 이뤄내고 있다. 하지만 신경망 구조의 내재적 특성 중 블랙박스 특성이 있는데 이는 모델의 예측 결과에 대한 해석이 어렵다는 특성을 의미한다. 이 블랙박스 특성은 인공지능 기술의 현장 적용을 방해하는데, 그 이유는 돌발적인 입력이 들어왔을 때 예측 결과를 사람이 이해할 수 없기 때문이다. 이러한 현상은 신뢰도 하락으로 이어지며, 예측 결과가 중대한 사안을 다루는 일일수록 인공지능 모델의 적용을 기피하게 되는 원인으로 작용한다.

[0004] 이러한 문제를 개선하기 위해 설명 가능한 인공지능(XAI: eXplainable Artificial Intelligence) 기술에 대한 연구개발이 활발히 진행되고 있다. XAI 기술이라 함은 인공지능 시스템의 의사결정 과정과 결과를 사람이 이해할 수 있는 수준의 형태로 설명을 제공하는 기술을 의미한다.

[0005] 하지만 복잡한 신경망 구조를 띄고 있는 인공지능 모델에 대해 설명하는 것은 쉽지 않은 일이며, 이를 간접적으로 설명할 수 있는 방법이 반사실 설명이다. 반사실 설명이란 인공지능 분류기의 판단에 반하는 판단, 즉 반대사실을 얻으려면 기존 입력에서 어떠한 것들을 변화시켜야 하는지 설명하는 것이다. 이는 주어진 입력과 모델의 예측 결과에 대해 "만약 이러한 조건이 달라졌다면, 결과는 어땠을까?"라는 가상의 상황을 가정하여 설명하는 방식이다.

[0006] 이러한 반사실 설명은 현실적(practical)이고 그럴듯한(plausible) 반사실을 생성하는 것을 목적으로 하나, 현실성과 그럴듯함의 특성은 서로 트레이드 오프(trade-off) 관계에 있기 때문에 이런 특성을 동시에 고려하여 반사실을 생성하는 것은 쉽지 않은 문제이다.

선행기술문헌

특허문헌

[0008] (특허문헌 0001) 한국공개특허공보 제10-2019-0078710호

발명의 내용

해결하려는 과제

[0009] 본 발명은 상기와 같은 문제를 해결하기 위해 안출된 것으로, 본 발명의 목적은 분류기의 판단에 대한 반사실 설명 생성 방법, 장치 및 기록 매체를 제공하는 것이다.

과제의 해결 수단

[0011] 상기 목적을 달성하기 위한 본 발명의 일 실시예에 따른 분류기의 판단에 대한 반사실 설명 생성 방법은, 입력 샘플과 분류기(classifier)에 대한 타겟 클래스 정보에 대한 참조 반사실 정보를 수신하는 단계; 상기 참조 반사실 정보로부터 상기 입력 샘플과 참조 반사실 정보에 대한 잠재 벡터를 도출하는 단계; 상기 입력 샘플의 잠재 벡터를 상기 참조 반사실 정보의 잠재 벡터에 대해 변환시키면서 상기 분류기의 클래스 변경에 대한 결정 경계(Decision Boundary)를 확인하는 단계; 및 상기 변환된 잠재 벡터가 적용된 입력 샘플이 상기 분류기의 결정 경계를 넘어서는 시점의 잠재 벡터를 검출하고, 검출된 잠재 벡터를 기반으로 반사실 설명을 생성하는 단계;를 포함한다.

[0012] 상기 목적을 달성하기 위한 본 발명의 일 실시예에 따른 분류기의 판단에 대한 반사실 설명 생성 장치는, 입력 샘플과 분류기(classifier)에 대한 타겟 클래스 정보에 대한 참조 반사실 정보를 수신하는 수신부; 상기 참조 반사실 정보로부터 상기 입력 샘플과 참조 반사실 정보에 대한 잠재 벡터를 도출하는 잠재 벡터 도출부; 상기 입력 샘플의 잠재 벡터를 상기 참조 반사실 정보의 잠재 벡터에 대해 변환시키면서 상기 분류기의 클래스 변경에 대한 결정 경계(Decision Boundary)를 확인하는 의미정보 변환부; 상기 변환된 잠재 벡터가 적용된 입력 샘플이 상기 분류기의 결정 경계를 넘어서는 시점의 잠재 벡터를 검출하고, 검출된 잠재 벡터를 기반으로 반사실 설명을 생성하는 반사실 설명 생성부;를 포함한다.

발명의 효과

[0014] 본 발명의 일 측면에 따르면, 본 발명은 분류기의 판단에 대한 반사실 설명 생성 방법, 장치 및 기록 매체를 제공함으로써, 모든 반사실 클래스에 대해서 설명 제공 모델인 GAN의 추가 학습 없이 추론 및 벡터 조작만으로 현실적이고 그럴듯한 반사실 설명을 생성할 수 있다.

도면의 간단한 설명

[0016] 도 1은 본 발명의 실시예에 따른 반사실 설명 생성을 위한 전체 프레임 워크를 나타낸 도면,
 도 2는 본 발명의 실시예에 따른 반사실 설명 생성 장치의 내부 블록을 나타낸 장치도,
 도 3은 본 발명의 실시예에 따른 조건부 GAN의 학습 과정을 나타낸 도면,
 도 4는 GAN의 잠재 공간에서 잠재 벡터 변환 과정을 나타낸 도면,
 그리고, 도 5는 본 발명의 실시예에 따른 반사실 설명 생성 장치의 동작을 나타낸 순서도.

발명을 실시하기 위한 구체적인 내용

[0017] 후술하는 본 발명에 대한 상세한 설명은, 본 발명이 실시될 수 있는 특정 실시예를 예시로서 도시하는 첨부 도면을 참조한다. 이들 실시예는 당업자가 본 발명을 실시할 수 있기에 충분하도록 상세히 설명된다. 본 발명의 다양한 실시예는 서로 다르지만 상호 배타적일 필요는 없음이 이해되어야 한다. 예를 들어, 여기에 기재되어 있는 특정 형상, 구조 및 특성은 일 실시예와 관련하여 본 발명의 정신 및 범위를 벗어나지 않으면서 다른 실시예로 구현될 수 있다. 또한, 각각의 개시된 실시예 내의 개별 구성요소의 위치 또는 배치는 본 발명의 정신 및 범

위를 벗어나지 않으면서 변경될 수 있음이 이해되어야 한다. 따라서, 후술하는 상세한 설명은 한정적인 의미로서 취하려는 것이 아니며, 본 발명의 범위는, 적절하게 설명된다면, 그 청구항들이 주장하는 것과 균등한 모든 범위와 더불어 첨부된 청구항에 의해서만 한정된다. 도면에서 유사한 참조부호는 여러 측면에 걸쳐서 동일하거나 유사한 기능을 지칭한다.

[0018] 본 발명에 따른 구성요소들은 물리적인 구분이 아니라 기능적인 구분에 의해서 정의되는 구성요소들로써 각각이 수행하는 기능들에 의해서 정의될 수 있다. 각각의 구성요소들은 하드웨어 또는 각각의 기능을 수행하는 프로그램 코드 및 프로세싱 유닛으로 구현될 수 있을 것이며, 두 개 이상의 구성요소의 기능이 하나의 구성요소에 포함되어 구현될 수도 있을 것이다. 따라서 이하의 실시예에서 구성요소에 부여되는 명칭은 각각의 구성요소를 물리적으로 구분하기 위한 것이 아니라 각각의 구성요소가 수행되는 대표적인 기능을 암시하기 위해서 부여된 것이며, 구성요소의 명칭에 의해서 본 발명의 기술적 사상이 한정되지 않는 것임에 유의하여야 한다.

[0019] 적대적 생성 신경망(GAN: Generative Adversarial Network)은 데이터 생성 능력에서 뛰어난 성능을 인정받고 있는 인공지능 모델이며, GAN의 엔코더와 디코더 사이의 잠재 공간(Latent space)을 활용하면 목적에 맞는 다양한 데이터를 생성할 수 있다. 특히, 잠재 공간에서 의미(semantic) 정보들이 얽히지 않고 풀려 있다면(disentangled) 잠재 벡터의 조작을 통해 의미 정보들을 선형적으로 변화시킬 수 있다는 장점이 있다. 따라서 본 발명의 실시예에서는 GAN의 의미 정보들이 얽히지 않고 풀려 있는 공간, 즉 잠재 공간에서 의미 정보를 선형적으로 변화시킬 수 있는 특성을 활용하여 현실적이고 그럴듯한 반사실 설명을 생성하는 방안에 대하여 설명하도록 한다.

[0020] 이하에서는 도면들을 참조하여 본 발명의 바람직한 실시예들을 보다 상세하게 설명하기로 한다.

[0021] 도 1은 본 발명의 실시예에 따른 반사실 설명 생성을 위한 전체 프레임워크를 나타낸 도면이고, 도 3은 본 발명의 실시예에 따른 조건부 GAN의 학습 과정을 나타낸 도면이다.

[0022] 입력 샘플과 분류기(classifier)에 대한 타겟 클래스 정보가 조건부 GAN(100)에 입력되고, 조건부 GAN(100)는 상기 입력 샘플과 상기 타겟 클래스 정보를 기반으로 생성된 참조 반사실 정보를 반사실 설명 생성 장치(200)로 출력한다.

[0023] 조건부 GAN(100)은 도 3에 도시한 바와 같은 학습 과정을 거쳐 학습되며, 그 상세 동작은 다음과 같다.

[0024] 조건부 GAN(100)은 생성자(G: Generator) 모델과 판별자(D: Discriminator) 모델로 구성되고, 상기 생성자 모델은 엔코더(ENC: Encoder), 디코더(DEC: Decoder) 및 상기 엔코더와 디코더 사이의 잠재 공간(W)을 포함한다.

[0025] 생성자 모델은 입력 샘플 x 와 타겟 클래스 정보 $\hat{y}$ 을 기반으로 참조 반사실 정보 $\hat{x}$ 을 생성하고, 판별자 모델은 생성자 모델이 생성한 참조 반사실 정보가 실제(Real) 입력 샘플인지 조작된(Fake) 샘플인지 판별한다.

[0026] 이때 상기 생성자 모델과 상기 판별자 모델은 비용 함수의 값이 최소화되는 방향으로 학습되며, 상기 생성자 모델의 비용 함수는 하기 수학적 식 1과 같이 나타낼 수 있고, 상기 판별자 모델의 비용 함수는 하기 수학적 식 2와 같이 나타낼 수 있다.

수학적 식 1

$$L_G = L_{adv} + \lambda_{cls} L_{cls}^f + \lambda_{rec} L_{rec}$$

[0027]

수학적 식 2

$$L_D = -L_{adv} + \lambda_{cls} L_{cls}^r$$

[0028]

[0029] 여기서 L_{adv} 는 실제 데이터 분포와 같은 샘플을 만들 수 있도록 도와주는 일반적인 적대적 손실을 나타내고, L_{cls}^f 는 생성된 샘플인 참조 반사실 정보가 타겟 클래스 정보에 따라 생성되도록 강제하는 상기 생성된 샘플에 대한 클래스 손실을 나타내고, L_{rec} 는 샘플 생성 시 입력 샘플과 비슷하게 생성되도록 강제하는 일관성 손실을 나타내고, λ_{cls} 는 각 손실 항의 기여도를 결정하는 계수를 나타낸다. 또한 L_{cls}^r 는 실제 입력 샘플의 클래스를 판별기가 구분할 수 있도록 하는 상기 실제 입력 샘플에 대한 클래스 손실을 나타내고, 이때 실제 입력 샘플의 클래스는 분류기에 의해 예측된 클래스를 사용한다.

[0030] 상기 일반적인 적대적 손실 L_{adv} 는 하기 수학식 3과 같이 나타낼 수 있고, 상기 생성된 샘플에 대한 클래스 손실 L_{cls}^f 는 하기 수학식 4와 같이 나타낼 수 있고, 상기 일관성 손실 L_{rec} 는 하기 수학식 5와 같이 나타낼 수 있고, 상기 실제 입력 샘플에 대한 클래스 손실 L_{cls}^r 은 하기 수학식 6과 같이 나타낼 수 있다.

수학식 3

$$L_{adv} = E_x [\log D_{src}(x)] + E_{x, \hat{y}} [\log (1 - D_{src}(G(x, \hat{y})))]$$

수학식 4

$$L_{cls}^f = E_{x, \hat{y}} [-\log D_{cls}(\hat{y} | G(x, \hat{y}))]$$

수학식 5

$$L_{rec} = E_{x, y, \hat{y}} [\|x - G(G(x, \hat{y}), y)\|_1]$$

수학식 6

$$L_{cls}^r = E_{x, y} [-\log D_{cls}(y | x)]$$

[0035] 여기서 E는 기대 값을 나타내고, x는 실제 입력 샘플을 나타내고, y는 상기 x의 클래스 정보를 나타내고, $\hat{x}$ 은 참조 반사실 정보를 나타내고, $\hat{y}$ 은 상기 $\hat{x}$ 의 클래스 정보를 나타낸다.

[0036] 한편, 반사실 설명 생성 장치(200)는 조건부 GAN(100)으로부터 입력된 참조 반사실 정보를 이용하여 반사실 설명을 생성하고, 생성된 반사실 설명을 출력한다. 반사실 설명 생성 장치(200)가 상기 반사실 설명을 생성하는 동작에 대해서는 후술할 도 2 및 도 4를 통해 보다 상세히 설명하도록 한다.

[0037] 도 2는 본 발명의 실시예에 따른 반사실 설명 생성 장치의 내부 블록을 나타낸 장치도이고, 도 4는 GAN의 잠재 공간에서 잠재 벡터 변환 과정을 나타낸 도면이다.

[0038] 반사실 설명 생성 장치(200)는 수신부(202), 잠재벡터 도출부(204), 의미정보 변환부(206) 및 반사실 설명 생성부(208)를 포함한다.

[0039] 수신부(202)는 입력 샘플과 분류기에 대한 타겟 클래스 정보에 대한 참조 반사실 정보를 수신한다. 여기서 상기 참조 반사실 정보는 미리 마련된 조건부 GAN에 상기 입력 샘플과 타겟 클래스 정보를 입력하여 획득된 결과를

의미한다.

[0040] 잠재벡터 도출부(204)는 상기 참조 반사실 정보로부터 상기 입력 샘플과 참조 반사실 정보 각각에 대한 잠재 벡터를 도출한다. 즉 잠재벡터 도출부(204)는 상기 참조 반사실 정보에 최적화 기반의 GAN 인버전(inversion)을 적용하여 상기 잠재 벡터를 도출한다. 이때 최적화하는 손실함수는 하기 수학적식 7과 같이 나타낼 수 있다.

수학적식 7

$$w^* = \arg \min_w \| DEC(w) - x \|_2^2 + \| ENC(DEC(w), \hat{y}) - ENC(x, \hat{y}) \|_2^2$$

[0042] 수학적식 7에서 첫번째 항은 잠재 공간(W)에서의 잠재 벡터를 디코딩했을 때 입력 샘플이 나오도록 하는 직접적인 손실항이며, 두번째 항은 첫번째 항에서 인코더까지 통과시켜 잠재 벡터에서도 서로 같아지도록 강제하는 손실항이다. 이렇게 두 가지 제약조건을 통해서 입력 샘플 x에 대한 잠재 벡터 w^* 를 얻을 수 있으며, 참조 반사실 정보 $\hat{x}$ 에 대한 잠재 벡터 $\hat{w}$ 의 경우 인코더에 입력 샘플과 참조 반사실 정보의 클래스 정보를 적용하여 ($ENC(x, \hat{y})$) 얻을 수 있다.

[0043] 의미정보 변환부(206)는 입력 샘플 x에 대한 잠재 벡터 w^* 를 상기 참조 반사실 정보 $\hat{x}$ 에 대한 잠재 벡터 $\hat{w}$ 에 대해 변환시키면서 분류기의 클래스 변경에 대한 결정 경계(Decision Boundary)를 확인한다. 즉 의미정보 변환부(206)는 잠재 벡터 w^* 를 잠재 벡터 $\hat{w}$ 방향으로 변환시키면서 상기 결정 경계를 확인한다. 여기서 잠재 벡터 w^* 를 잠재 벡터 $\hat{w}$ 방향으로 변환시키는 것은, 잠재 벡터 w^* 와 잠재 벡터 $\hat{w}$ 사이에 선형 내삽을 진행하여 도 4에 도시한 바와 같이 선형적으로 벡터 값을 변환하는 것을 의미한다. 이때 상기 선형 내삽은 하기 수학적식 8을 통해 진행된다.

수학적식 8

$$\hat{w}^* = (1 - \lambda)w^* + \lambda\hat{w}, \lambda \in (0, 1)$$

[0045] 이와 같이 의미정보 변환부(206)는 잠재 벡터의 선형 변환을 통해 의미 정보인 클래스 정보만을 변환시키면서 분류기의 결정 경계를 넘어서는지 여부를 확인할 수 있다.

[0046] 한편, 잠재벡터 도출부(204)와 의미정보 변환부(206)는 인코더와 디코더 사이의 잠재 공간 W를 활용하여 동작하고, 여기서 잠재 공간은 의미 정보들이 얹히지 않고 풀린 공간을 의미한다. 잠재 공간에서 의미 정보들이 풀려 있는 이유는 GAN의 학습 과정에서 인코더를 통해 압축된 정보로부터 디코딩을 통해 샘플을 생성할 때 좀더 실제 데이터와 같은 샘플을 쉽게 생성하기 위해서 의미 정보들이 잘 맵핑 되도록 하기 위함이다. 수학적으로 보면 이 공간은 Z 공간처럼 가우시안 구체 분포 형태와 같이 특정 형태를 띄어야 한다는 제약 조건 또한 존재하지 않기 때문에 앞서 언급한 특징이 더 부각될 수 있다.

[0047] 반사실 설명 생성부(208)는 의미정보 변환부(206)의 확인 결과에 따라 변환된 잠재 벡터가 적용된 입력 샘플이 분류기의 결정 경계를 넘어서는 시점의 잠재 벡터를 검출하고, 검출된 잠재 벡터를 기반으로 반사실 설명을 생성한다. 즉 반사실 설명 생성부(208)는 검출된 잠재 벡터를 디코딩하여 상기 반사실 설명을 생성한다. 이렇게 생성된 반사실 설명은 입력으로부터 의미 정보의 변화를 최소화 하였기 때문에 사용자로 하여금 현실화가 용이한 반사실이며, 의미 정보만을 적절히 변화시켰기 때문에 반사실 클래스 데이터처럼 그럴듯한 반사실이 된다.

[0048] 도 5는 본 발명의 실시예에 따른 반사실 설명 생성 장치의 동작을 나타낸 순서도이다.

[0049] 반사실 설명 생성 장치는 입력 샘플과 분류기에 대한 타겟 클래스 정보를 조건부 GAN에 입력하여 획득된 참조 반사실 정보를 수신하고, (S502) 수신된 참조 반사실 정보에 GAN 인버전(inversion)을 적용하여 상기 입력 샘플 및 참조 반사실 정보 각각에 대한 잠재 벡터를 도출한다. (S504)

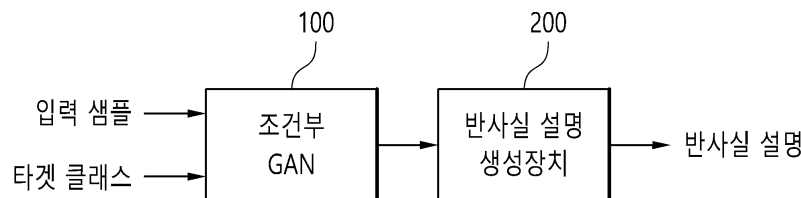
[0050] 반사실 설명 생성 장치는 상기 도출된 입력 샘플에 대한 잠재 벡터를 상기 참조 반사실 정보에 대한 잠재 벡터

방향으로 변환시키면서 분류기의 클래스 변경에 대한 결정 경계를 확인한다.(S506)

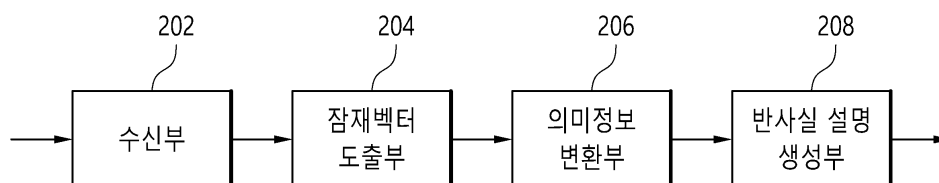
- [0051] 이후 반사실 설명 생성 장치는 S506에서 변환된 잠재 벡터가 적용된 입력 샘플이 상기 분류기의 결정 경계를 넘어서는 시점의 잠재 벡터를 검출하고, 검출된 잠재 벡터를 기반으로 반사실 설명을 생성한다.(S508)
- [0052] 이와 같은 본 발명의 분류기의 판단에 대한 반사실 설명 생성 방법은 다양한 컴퓨터 구성요소를 통하여 수행될 수 있는 프로그램 명령어의 형태로 구현되어 컴퓨터 판독 가능한 기록 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능한 기록 매체는 프로그램 명령어, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다.
- [0053] 상기 컴퓨터 판독 가능한 기록 매체에 기록되는 프로그램 명령어는 본 발명을 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 분야의 당업자에게 공지되어 사용 가능한 것일 수도 있다.
- [0054] 컴퓨터 판독 가능한 기록 매체의 예에는, 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체, CD-ROM, DVD 와 같은 광기록 매체, 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 ROM, RAM, 플래시 메모리 등과 같은 프로그램 명령어를 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다.
- [0055] 프로그램 명령어의 예에는, 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드도 포함된다. 상기 하드웨어 장치는 본 발명에 따른 처리를 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.
- [0056] 이상에서는 본 발명의 다양한 실시예에 대하여 도시하고 설명하였지만, 본 발명은 상술한 특징의 실시예에 한정되지 아니하며, 청구범위에서 청구하는 본 발명의 요지를 벗어남이 없이 당해 발명이 속하는 기술분야에서 통상의 지식을 가진 자에 의해 다양한 변형실시가 가능한 것은 물론이고, 이러한 변형실시들은 본 발명의 기술적 사상이나 전망으로부터 개별적으로 이해되어져서는 안될 것이다.

도면

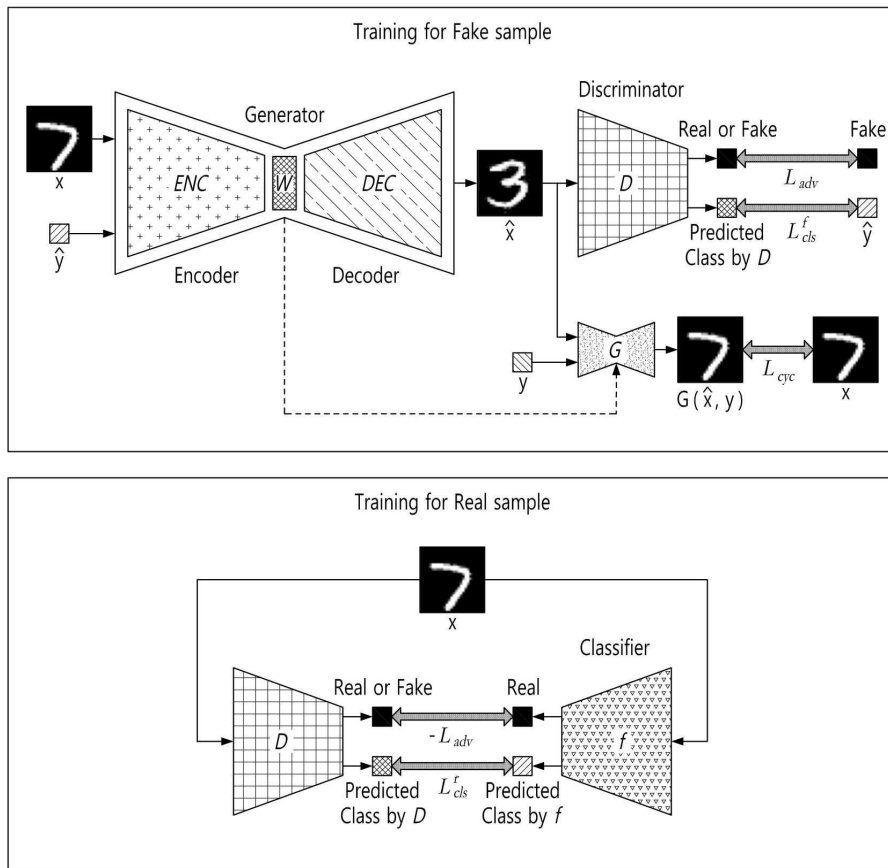
도면1



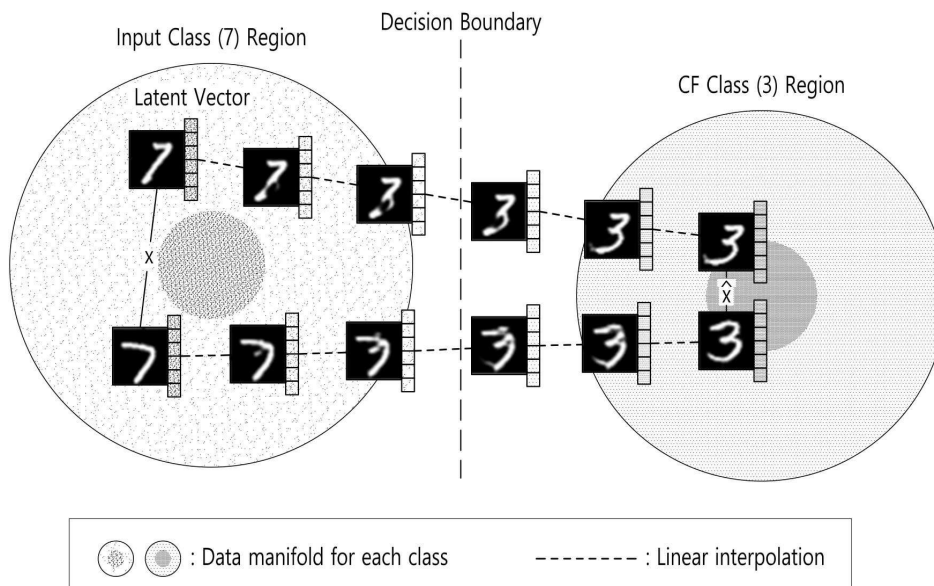
도면2



도면3



도면4



도면5

