



(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2023년10월24일
(11) 등록번호 10-2593137
(24) 등록일자 2023년10월19일

- (51) 국제특허분류(Int. Cl.)
G06V 10/764 (2022.01) G06N 3/08 (2023.01)
G06V 10/46 (2022.01) G06V 10/774 (2022.01)
G06V 10/80 (2022.01) G06V 30/182 (2022.01)
G06V 30/19 (2022.01)
- (52) CPC특허분류
G06V 10/764 (2023.08)
G06N 3/08 (2023.01)
- (21) 출원번호 10-2023-0049560
(22) 출원일자 2023년04월14일
심사청구일자 2023년04월14일
- (56) 선행기술조사문헌
KR1020210065390 A*
(뒷면에 계속)
- (73) 특허권자
고려대학교산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
- (72) 발명자
김진규
서울시 광진구 아차산로 599 신동아파밀리에 107동 801호
정유진
서울특별시 성북구 고려대로 45 (안암동2가) 파인빌동 305호
박성범
경기도 화성시봉담읍 동화새터길 55-39 (동화마을 신동아파밀리에아파트) 108동 1301호
- (74) 대리인
특허법인위솔

전체 청구항 수 : 총 8 항

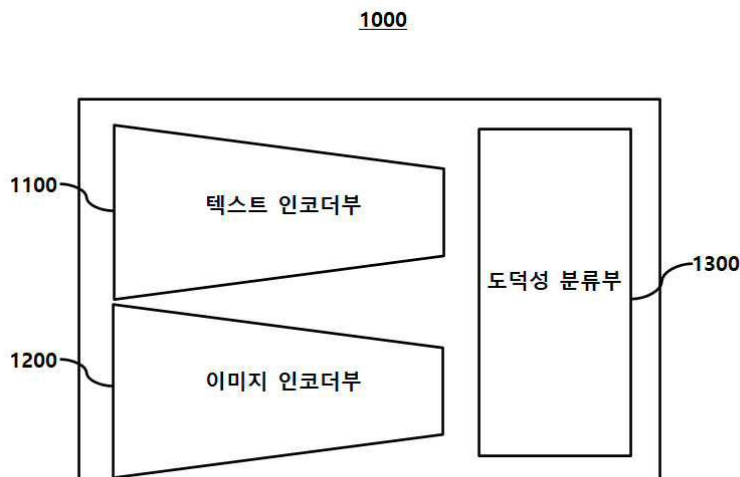
심사관 : 강현일

(54) 발명의 명칭 딥러닝 기술을 이용한 비도덕적인 이미지 분류 장치 및 방법

(57) 요약

본 발명의 일 실시 예에 따른 비도덕적인 이미지 분류 장치는 텍스트를 입력 받아 이로부터 텍스트적 임베딩 벡터(Textual Embedding Vector)를 생성하는 텍스트 인코더부, 이미지를 입력 받아 이로부터 시각적 임베딩 벡터(Visual Embedding Vector)를 생성하는 이미지 인코더부 및 상기 텍스트적 임베딩 벡터 또는 시각적 임베딩 벡터 중 어느 하나를 입력 받아 상기 입력 받은 텍스트 또는 이미지가 도덕적(Moral)인지 비도덕적(Immoral)인지의 분류 결과를 생성하여 출력하는 도덕성 분류부를 포함하며, 상기 도덕성 분류부는, 도덕적 또는 비도덕적인지 여부가 이진 클래스로 매핑되어 있는 학습용 텍스트를 복수 개 포함하는 학습용 데이터셋만으로 입력 받은 학습용 텍스트의 분류 결과를 학습한다.

대표도 - 도2



(52) CPC특허분류

G06V 10/469 (2023.08)
G06V 10/774 (2023.08)
G06V 10/803 (2023.08)
G06V 30/1823 (2023.08)
G06V 30/19147 (2023.08)

(56) 선행기술조사문헌

KR102410582 B1
KR102358186 B1
KR1020230029281 A
KR1020220122385 A
KR1020230018952 A

*는 심사관에 의하여 인용된 문헌

이 발명을 지원한 국가연구개발사업

과제고유번호	1711193454
과제번호	2022-0-00043-002
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	사람중심인공지능핵심원천기술개발
연구과제명	개성 형성이 가능한 에이전트 플랫폼 기술 개발
기 여 율	1/2
과제수행기관명	한국전자기술연구원
연구기간	2023.01.01 ~ 2023.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1711193663
과제번호	2020-0-01819-004
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신방송혁신인재양성
연구과제명	ICT명품인재양성(고려대학교)
기 여 율	1/2
과제수행기관명	고려대학교산학협력단
연구기간	2023.01.01 ~ 2023.12.31

공지예외적용 : 있음

명세서

청구범위

청구항 1

텍스트를 입력 받아 이로부터 텍스트적 임베딩 벡터(Textual Embedding Vector)를 생성하는 텍스트 인코더부;
 이미지를 입력 받아 이로부터 시각적 임베딩 벡터(Visual Embedding Vector)를 생성하는 이미지 인코더부; 및
 상기 텍스트적 임베딩 벡터 또는 시각적 임베딩 벡터 중 어느 하나를 입력 받아 상기 입력 받은 텍스트 또는 이미지가 도덕적(Moral)인지 비도덕적(Immoral)인지의 분류 결과를 생성하여 출력하는 도덕성 분류부;
 를 포함하며,
 상기 도덕성 분류부는,
 도덕적 또는 비도덕적인지 여부가 이진 클래스로 맵핑되어 있는 학습용 텍스트를 복수 개 포함하는 학습용 데이터셋만으로 입력 받은 학습용 텍스트의 분류 결과를 학습하며,
 복수 개의 Fully Connected Layer를 포함하는 MLP(Multi-Layer Perceptron)으로 이루어지고,
 상기 수학적식에 따른 손실 함수(Lc)를 이용하여 학습을 진행하는,
 비도덕적인 이미지 분류 장치.

$$\text{손실 함수}(L_c) = -\frac{1}{n} \sum_{i=1}^n \log \sigma(\hat{y}_i) + (-y_i) \log(1 - \sigma(\hat{y}_i))]$$

여기서 n은 minibatch 내부의 텍스트의 개수, y는 정답 라벨만 1이고 나머지는 0인 one-hot 벡터, $\hat{y}$ 는 상기 장치가 텍스트적 임베딩 벡터를 입력 받아 출력한 분류 결과, σ 는 Sigmoid Function을 의미함

청구항 2

제1항에 있어서,
 상기 텍스트 인코더부는,
 텍스트 및 이에 대응되는 이미지 한 쌍을 동일한 임베딩 공간(Embedding Space)에 맵핑하는 사전 학습된 CLIP(Contrastive Language-Image Pre-training) 모델 기반 텍스트 인코더인,
 비도덕적인 이미지 분류 장치.

청구항 3

제2항에 있어서,
 상기 텍스트적 임베딩 벡터는,
 상기 입력 받은 텍스트의 특징자 또는 특징 벡터 및 이에 대응되는 이미지의 특징자 또는 특징 벡터가 매칭되어 맵핑된 조인트 임베딩 공간의 임베딩 벡터인,
 비도덕적인 이미지 분류 장치.

청구항 4

제1항에 있어서,

상기 이미지 인코더부는,

텍스트 및 이에 대응되는 이미지 한 쌍을 동일한 임베딩 공간(Embedding Space)에 맵핑하는 사전 학습된 CLIP(Contrastive Language-Image Pre-training) 모델 기반 이미지 인코더인,

비도덕적인 이미지 분류 장치.

청구항 5

제4항에 있어서,

상기 시각적 임베딩 벡터는,

상기 입력 받은 이미지의 특징자 또는 특징 벡터 및 이에 대응되는 텍스트의 특징자 또는 특징 벡터가 매칭되어 맵핑된 조인트 임베딩 공간의 임베딩 벡터인,

비도덕적인 이미지 분류 장치.

청구항 6

삭제

청구항 7

제1항에 있어서,

상기 복수 개의 Fully Connected Layer는,

Dropout Layer - Linear Layer - Tanh Layer - Dropout Layer - Projection Layer인,

비도덕적인 이미지 분류 장치.

청구항 8

삭제

청구항 9

프로세서 및 메모리를 포함하는 장치가 비도덕적인 이미지를 분류하는 방법에 있어서,

(a) 학습용 데이터셋을 입력 받아 이로부터 텍스트적 임베딩 벡터(Textual Embedding Vector)를 생성하고, 상기 생성한 텍스트적 임베딩 벡터를 입력 받아 학습용 텍스트가 도덕적인지 비도덕적인지 분류 결과를 출력하되, 손실 함수(Lc)를 최소화하는 방향으로 학습을 수행하는 제1 단계; 및

(b) 실행용 이미지를 입력 받아 상기 실행용 이미지가 도덕적(Moral)인지 비도덕적(Immoral)인지의 분류 결과를 출력하는 제2 단계;

를 포함하며,

상기 학습용 데이터셋은,

도덕적 또는 비도덕적인지 여부가 이진 클래스로 맵핑되어 있는 학습용 텍스트만을 복수 개 포함하고,

상기 제1 단계에서 분류 결과의 출력은,

복수 개의 Fully Connected Layer를 포함하는 MLP(Multi-Layer Perceptron)를 통과하여 이루어지며,

상기 손실 함수(Lc)는 하기와 같은 비도덕적인 이미지 분류 방법.

$$\text{손실 함수(Lc)} = -\frac{1}{n} \sum_{i=1}^n \log \sigma(\hat{y}_i) + (-y_i) \log (1 - \sigma(\hat{y}_i))]$$

여기서 n은 minibatch 내부의 텍스트의 개수, y는 정답 라벨만 1이고 나머지는 0인 one-hot 벡터, $\hat{y}$ 는 상기 장치가 텍스트적 임베딩 벡터를 입력 받아 출력한 분류 결과, σ 는 Sigmoid Function을 의미함

청구항 10

컴퓨팅 장치와 결합하여,

(AA) 학습용 데이터셋을 입력 받아 이로부터 텍스트적 임베딩 벡터(Textual Embedding Vector)를 생성하고, 상기 생성한 텍스트적 임베딩 벡터를 입력 받아 학습용 텍스트가 도덕적인지 비도덕적인지 분류 결과를 출력하되, 손실 함수(Lc)를 최소화하는 방향으로 학습을 수행하는 제1 단계; 및

(BB) 실행용 이미지를 입력 받아 상기 실행용 이미지가 도덕적(Moral)인지 비도덕적(Immoral)인지의 분류 결과를 출력하는 제2 단계;

를 실행하며,

상기 학습용 데이터셋은,

도덕적 또는 비도덕적인지 여부가 이진 클래스로 맵핑되어 있는 학습용 텍스트만을 복수 개 포함하고,

상기 제1 단계에서 분류 결과의 출력은,

복수 개의 Fully Connected Layer를 포함하는 MLP(Multi-Layer Perceptron)를 통과하여 이루어지며,

상기 손실 함수(Lc)는 하기와 같은 컴퓨터로 판독 가능한 매체에 저장된 컴퓨터 프로그램.

$$\text{손실 함수(Lc)} = -\frac{1}{n} \sum_{i=1}^n \log \sigma(\hat{y}_i) + (-y_i) \log (1 - \sigma(\hat{y}_i))]$$

여기서 n은 minibatch 내부의 텍스트의 개수, y는 정답 라벨만 1이고 나머지는 0인 one-hot 벡터, $\hat{y}$ 는 상기 컴퓨터 프로그램이 텍스트적 임베딩 벡터를 입력 받아 출력한 분류 결과, σ 는 Sigmoid Function을 의미함

발명의 설명

기술 분야

[0001] 본 발명은 비도덕적인 이미지 분류 장치 및 방법에 관한 것이다. 보다 자세하게는 이미지에 대한 학습을 진행하지 않고도 비도덕적인 이미지를 분류할 수 있는 장치 및 방법에 관한 것이다.

배경 기술

[0003] 2020년 전세계 디지털 콘텐츠 시장이 2조 5,160억 달러를 기록했으며, 향후 디지털 콘텐츠 시장은 네트워크 고도화와 모바일 라이프의 확산 추세를 타고 그 규모가 더욱 커질 것으로 전망되고 있는바, 2025년까지 연평균 12.0% 성장한 4조 4,270억에 이를 것으로 내다보고 있다.

[0004] 이와 같이 디지털 콘텐츠에 대한 수요가 국/내외적으로 증가함에 따라 다양한 기관 및 기업들에서 비도덕적인 콘텐츠를 자동으로 분류할 수 있는 기술 개발에 박차를 가하고 있는바, 종래의 비도덕적인 콘텐츠 분류 기술은 대부분 그 대상을 텍스트인 자연어에 치중하고 있었다.

[0005] 또한, 일부 비도덕적인 이미지의 분류 기술은 폭력성, 선정성 등과 같이 특정 분야에만 치중되어 해당 분야에 관한 특정 목적에 제한적으로만 적용 가능한 경우가 대부분인바, 특정 분야에 치중되지 않은 인간의 보편적인 사고 수준에서 비도덕적이라고 인정될만한 이미지를 분류하는 것이 매우 어렵다는 문제점이 있으며, 인공지능 모델에 이를 학습시켜 분류하고자 한다 할지라도 인간의 보편적인 도덕성을 지닌 적절한 시각적 학습 데이터셋이 존재하지 않기에 마찬가지로 어려움이 있는바, 보다 근본적인 해결 방안이 요구된다. 본 발명은 이에 관한

것이다.

선행기술문헌

특허문헌

[0007] (특허문헌 0001) 대한민국 공개특허공보 제 10-2020-0163934호(2022.06.08)

발명의 내용

해결하려는 과제

[0008] 본 발명이 해결하고자 하는 기술적 과제는 특정 분야에 치중되지 않은 인간의 보편적인 사고 수준에서 비도덕적이라고 인정될만한 이미지를 분류할 수 있는 비도덕적인 이미지 분류 장치 및 방법을 제공하는 것이다.

[0009] 본 발명이 해결하고자 하는 또 다른 기술적 과제는 인간의 보편적인 도덕성을 지닌 적절한 시각적 학습 데이터셋이 존재하지 않는 상황 하에서도 비도덕적인 이미지를 효과적으로 분류할 수 있는 비도덕적인 이미지 분류 장치 및 방법을 제공하는 것이다.

[0010] 본 발명의 기술적 과제들은 이상에서 언급한 기술적 과제들로 제한되지 않으며, 언급되지 않은 또 다른 기술적 과제들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

과제의 해결 수단

[0012] 상기 기술적 과제를 달성하기 위한 본 발명의 일 실시 예에 따른 비도덕적인 이미지 분류 장치는 텍스트를 입력받아 이로부터 텍스트적 임베딩 벡터(Textual Embedding Vector)를 생성하는 텍스트 인코더부, 이미지를 입력받아 이로부터 시각적 임베딩 벡터(Visual Embedding Vector)를 생성하는 이미지 인코더부 및 상기 텍스트적 임베딩 벡터 또는 시각적 임베딩 벡터 중 어느 하나를 입력받아 상기 입력 받은 텍스트 또는 이미지가 도덕적(Moral)인지 비도덕적(Immoral)인지의 분류 결과를 생성하여 출력하는 도덕성 분류부를 포함하며, 상기 도덕성 분류부는, 도덕적 또는 비도덕적인지 여부가 이진 클래스로 맵핑되어 있는 학습용 텍스트를 복수 개 포함하는 학습용 데이터셋만으로 입력 받은 학습용 텍스트의 분류 결과를 학습한다.

[0013] 일 실시 예에 따르면, 상기 텍스트 인코더부는, 텍스트 및 이에 대응되는 이미지 한 쌍을 동일한 임베딩 공간(Embedding Space)에 맵핑하는 사전 학습된 CLIP(Contrastive Language-Image Pre-training) 모델 기반 텍스트 인코더일 수 있다.

[0014] 일 실시 예에 따르면, 상기 텍스트적 임베딩 벡터는, 상기 입력 받은 텍스트의 특징자 또는 특징 벡터 및 이에 대응되는 이미지의 특징자 또는 특징 벡터가 매칭되어 맵핑된 조인트 임베딩 공간의 임베딩 벡터일 수 있다.

[0015] 일 실시 예에 따르면, 상기 이미지 인코더부는, 텍스트 및 이에 대응되는 이미지 한 쌍을 동일한 임베딩 공간(Embedding Space)에 맵핑하는 사전 학습된 CLIP(Contrastive Language-Image Pre-training) 모델 기반 이미지 인코더일 수 있다.

[0016] 일 실시 예에 따르면, 상기 시각적 임베딩 벡터는, 상기 입력 받은 이미지의 특징자 또는 특징 벡터 및 이에 대응되는 텍스트의 특징자 또는 특징 벡터가 매칭되어 맵핑된 조인트 임베딩 공간의 임베딩 벡터일 수 있다.

[0017] 일 실시 예에 따르면, 상기 도덕성 분류부는, 복수 개의 Fully Connected Layer를 포함하는 MLP(Multi-Layer Perceptron)으로 이루어질 수 있다.

[0018] 일 실시 예에 따르면, 상기 복수 개의 Fully Connected Layer는, Dropout Layer - Linear Layer - Tanh Layer - Dropout Layer - Projection Layer일 수 있다.

[0019] 일 실시 예에 따르면, 상기 도덕성 분류부는, 하기 수학적식에 따른 손실 함수(Lc)를 이용하여 학습을 진행할 수 있다.

[0020] 손실 함수(Lc) =
$$-\frac{1}{n} \sum_{i=1}^n \log \sigma(\hat{y}_i) + (-y_i) \log(1 - \sigma(\hat{y}_i))$$
]] Function을 의미함

[0021] 여기서 n 은 minibatch 내부의 텍스트의 개수, y 는 정답 라벨만 1이고 나머지는 0인 one-hot 벡터, $\hat{y}$ 는 도덕성 분류부가 텍스트적 임베딩 벡터를 입력 받아 출력한 분류 결과, σ 는 Sigmoid Function을 의미함

[0022] 상기 기술적 과제를 달성하기 위한 본 발명의 또 다른 실시 예에 따른 프로세서 및 메모리를 포함하는 장치가 비도덕적인 이미지 분류 방법은 (a) 학습용 데이터셋을 입력 받아 이로부터 텍스트적 임베딩 벡터(Textual Embedding Vector)를 생성하고, 상기 생성한 텍스트적 임베딩 벡터를 입력 받아 상기 학습용 텍스트가 도덕적인지 비도덕적인지 분류 결과를 출력하되, 손실 함수(Lc)를 최소화하는 방향으로 학습을 수행하는 제1 단계 및 (b) 실행용 이미지를 입력 받아 상기 실행용 이미지가 도덕적(Moral)인지 비도덕적(Immoral)인지의 분류 결과를 출력하는 제2 단계를 포함하며, 상기 학습용 데이터셋은, 도덕적 또는 비도덕적인지 여부가 이진 클래스로 맵핑되어 있는 학습용 텍스트만을 복수 개 포함하고, 상기 손실 함수(Lc)는 하기와 같다.

[0023] 손실 함수(Lc) =
$$-\frac{1}{n} \sum_{i=1}^n \log \sigma(\hat{y}_i) + (-y_i) \log (1 - \sigma(\hat{y}_i))$$

[0024] 여기서 n 은 minibatch 내부의 텍스트의 개수, y 는 정답 라벨만 1이고 나머지는 0인 one-hot 벡터, $\hat{y}$ 는 상기 장치가 텍스트적 임베딩 벡터를 입력 받아 출력한 분류 결과, σ 는 Sigmoid Function을 의미함

[0025] 상기 기술적 과제를 달성하기 위한 본 발명의 또 다른 실시 예에 따른 매체에 저장된 컴퓨터 프로그램은 컴퓨팅 장치와 결합하여, (AA) 학습용 데이터셋을 입력 받아 이로부터 텍스트적 임베딩 벡터(Textual Embedding Vector)를 생성하고, 상기 생성한 텍스트적 임베딩 벡터를 입력 받아 상기 학습용 텍스트가 도덕적인지 비도덕적인지 분류 결과를 출력하되, 손실 함수(Lc)를 최소화하는 방향으로 학습을 수행하는 제1 단계 및 (BB) 실행용 이미지를 입력 받아 상기 실행용 이미지가 도덕적(Moral)인지 비도덕적(Immoral)인지의 분류 결과를 출력하는 제2 단계를 실행하며, 상기 학습용 데이터셋은, 도덕적 또는 비도덕적인지 여부가 이진 클래스로 맵핑되어 있는 학습용 텍스트만을 복수 개 포함하고, 상기 손실 함수(Lc)는 하기와 같다.

[0026] 손실 함수(Lc) =
$$-\frac{1}{n} \sum_{i=1}^n \log \sigma(\hat{y}_i) + (-y_i) \log (1 - \sigma(\hat{y}_i))$$

[0027] 여기서 n 은 minibatch 내부의 텍스트의 개수, y 는 정답 라벨만 1이고 나머지는 0인 one-hot 벡터, $\hat{y}$ 는 상기 컴퓨터 프로그램이 텍스트적 임베딩 벡터를 입력 받아 출력한 분류 결과, σ 는 Sigmoid Function을 의미함

발명의 효과

[0029] 상기와 같은 본 발명에 따르면 인간의 보편적인 사고 수준에서 도덕적 또는 비도덕적인지 여부가 이진 클래스로 맵핑되어 있는 학습용 텍스트의 특징자 또는 특징 벡터가 이에 대응되는 이미지의 특징자 또는 특징 벡터와 매칭되어 맵핑된 조인트 임베딩 공간의 임베딩 벡터를 이용하여 도덕적인지 여부에 대한 분류 결과를 생성하여 출력하는 과정을 학습하므로, 학습 이후 어떠한 실행용 이미지가 입력된다 할지라도 해당 실행용 이미지의 특징자 또는 특징 벡터가 속하게 되는 조인트 임베딩 공간은 학습 과정에서 경험한 조인트 임베딩 공간일 것인바, 인간의 보편적인 도덕성을 지닌 적절한 시각적 학습 데이터셋이 존재하지 않는 상황 하에서도 비도덕적인 이미지를 효과적으로 분류할 수 있다.

[0030] 본 발명의 효과들은 이상에서 언급한 효과들로 제한되지 않으며, 언급되지 않은 또 다른 효과들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해 될 수 있을 것이다.

도면의 간단한 설명

[0032] 도 1은 본 발명의 제1 실시 예에 따른 비도덕적인 이미지 분류 장치가 포함하는 전체 구성을 예시적으로 도시한 도면이다.

도 2는 본 발명의 제2 실시 예에 따른 비도덕적인 이미지 분류 장치가 이미지 분류를 위해 수행하는 동작 또는 연산을 기능적 구성으로 명명하여 도시한 내부 구조도이다.

도 3은 ETHICS의 Commonsense Morality 데이터셋을 예시적으로 도시한 도면이다.

도 4는 CLIP 모델의 학습 모식도를 예시적으로 도시한 도면이다.

도 5는 본 발명의 제2 실시 예에 따른 비도덕적인 이미지 분류 장치의 학습 모식도를 예시적으로 도시한 도면이

다.

도 6은 본 발명의 제2 실시 예에 따른 비도덕적인 이미지 분류 장치의 샘플링 모식도를 예시적으로 도시한 도면이다.

도 7은 도 5 및 도 6을 실제 텍스트와 이미지를 이용해 병합하여 도시한 도면이다.

도 8 및 도 9는 본 발명의 제2 실시 예에 따른 비도덕적인 이미지 분류 장치의 성능 평가 자료이다.

도 10은 본 발명의 제3 실시 예에 따른 비도덕적인 이미지를 분류하는 방법의 대표적인 단계를 나타낸 순서도이다.

발명을 실시하기 위한 구체적인 내용

- [0033] 본 발명의 목적과 기술적 구성 및 그에 따른 작용 효과에 관한 자세한 사항은 본 발명의 명세서에 첨부된 도면에 의거한 이하의 상세한 설명에 의해 보다 명확하게 이해될 것이다. 첨부된 도면을 참조하여 본 발명에 따른 실시 예를 상세하게 설명한다.
- [0034] 본 명세서에서 개시되는 실시 예들은 본 발명의 범위를 한정하는 것으로 해석되거나 이용되지 않아야 할 것이다. 이 분야의 통상의 기술자에게 본 명세서의 실시 예를 포함한 설명은 다양한 응용을 갖는다는 것이 당연하다. 따라서, 본 발명의 상세한 설명에 기재된 임의의 실시 예들은 본 발명을 보다 잘 설명하기 위한 예시적인 것이며 본 발명의 범위가 실시 예들로 한정되는 것을 의도하지 않는다.
- [0035] 도면에 표시되고 아래에 설명되는 기능 블록들은 가능한 구현의 예들일 뿐이다. 다른 구현들에서는 상세한 설명의 사상 및 범위를 벗어나지 않는 범위에서 다른 기능 블록들이 사용될 수 있다. 또한, 본 발명의 하나 이상의 기능 블록이 개별 블록들로 표시되지만, 본 발명의 기능 블록들 중 하나 이상은 동일 기능을 실행하는 다양한 하드웨어 및 소프트웨어 구성들의 조합일 수 있다.
- [0036] 또한, 어떤 구성요소들을 포함한다는 표현은 "개방형"의 표현으로서 해당 구성요소들이 존재하는 것을 단순히 지칭할 뿐이며, 추가적인 구성요소들을 배제하는 것으로 이해되어서는 안 된다.
- [0037] 나아가 어떤 구성요소가 다른 구성요소에 "연결되어" 있다거나 "접속되어" 있다고 언급될 때에는, 그 다른 구성요소에 직접적으로 연결 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 한다.
- [0038] 이하에서는 도면들을 참조하여 본 발명의 세부적인 실시 예들에 대해 살펴보도록 한다.
- [0039] 도 1은 본 발명의 제1 실시 예에 따른 비도덕적인 이미지 분류 장치(100)가 포함하는 전체 구성을 예시적으로 도시한 도면이다.
- [0040] 그러나 이는 본 발명의 목적을 달성하기 위한 바람직한 실시 예일 뿐이며, 필요에 따라 일부 구성이 추가되거나 삭제될 수 있고, 어느 한 구성이 수행하는 역할을 다른 구성이 함께 수행할 수도 있음은 물론이다.
- [0041] 본 발명의 제1 실시 예에 따른 비도덕적인 이미지 분류 장치(100)는 프로세서(10), 네트워크 인터페이스(20), 메모리(30), 스토리지(40) 및 이들을 연결하는 데이터 버스(50)를 포함할 수 있으며, 기타 본 발명의 목적을 달성함에 있어 요구되는 부가적인 구성들을 더 포함할 수 있음은 물론이라 할 것이다.
- [0042] 프로세서(10)는 각 구성의 전반적인 동작을 제어한다. 프로세서(10)는 CPU(Central Processing Unit), MPU(Micro Processor Unit), MCU(Micro Controller Unit) 또는 본 발명이 속하는 기술 분야에서 널리 알려져 있는 형태의 인공지능 프로세서 중 어느 하나일 수 있다. 아울러, 프로세서(10)는 후술할 본 발명의 제2 실시 예에 따른 비도덕적인 이미지 분류 장치(1000)에서 설명할 각각의 동작을 수행하기 위한 적어도 하나의 애플리케이션 또는 프로그램에 대한 연산을 수행할 수 있다.
- [0043] 네트워크 인터페이스(20)는 본 발명의 제1 실시 예에 따른 비도덕적인 이미지 분류 장치(100)의 유무선 인터넷 통신을 지원하며, 그 밖의 공지의 통신 방식을 지원할 수도 있다. 따라서 네트워크 인터페이스(20)는 그에 따른 통신 모듈을 포함하여 구성될 수 있다.
- [0044] 메모리(30)는 각종 정보, 명령 및/또는 정보를 저장하며, 본 발명의 제3 실시 예에 따른 비도덕적인 이미지 분류 방법을 수행하기 위해 스토리지(40)로부터 하나 이상의 컴퓨터 프로그램(41)을 로드할 수 있다. 도 1에서는 메모리(30)의 하나로 RAM을 도시하였으나 이와 더불어 다양한 저장 매체를 메모리(30)로 이용할 수 있음은 물론

이다.

- [0045] 스토리지(40)는 하나 이상의 컴퓨터 프로그램(41) 및 대용량 네트워크 정보(42)를 비임시적으로 저장할 수 있다. 이러한 스토리지(40)는 ROM(Read Only Memory), EPROM(Erasable Programmable ROM), EEPROM(Electrically Erasable Programmable ROM), 플래시 메모리 등과 같은 비휘발성 메모리, 하드 디스크(HDD), 보조 저장 매체(SSD), 착탈형 디스크, 또는 본 발명이 속하는 기술 분야에서 널리 알려져 있는 임의의 형태의 컴퓨터로 읽을 수 있는 기록 매체 중 어느 하나일 수 있다.
- [0046] 컴퓨터 프로그램(41)은 메모리(30)에 로드되어, 하나 이상의 프로세서(10)에 의해, (A) 학습용 데이터셋을 입력 받아 이로부터 텍스트적 임베딩 벡터(Textual Embedding Vector)를 생성하고, 상기 생성한 텍스트적 임베딩 벡터를 입력 받아 상기 학습용 텍스트가 도덕적인지 비도덕적인지 분류 결과를 출력하되, 손실 함수(Lc)를 최소화하는 방향으로 학습을 수행하는 제1 오퍼레이션 및 (B) 실행용 이미지를 입력 받아 상기 실행용 이미지가 도덕적(Moral)인지 비도덕적(Immoral)인지의 분류 결과를 출력하는 제2 단계를 실행할 수 있다.
- [0047] 이상 간단하게 언급한 컴퓨터 프로그램(41)이 수행하는 오퍼레이션은 컴퓨터 프로그램(41)의 일 기능으로 볼 수 있으며, 보다 자세한 설명은 본 발명의 제2 실시 예에 따른 비도덕적인 이미지 분류 장치(1000)에 대한 설명에서 후술하도록 한다.
- [0048] 데이터 버스(50)는 이상 설명한 프로세서(10), 네트워크 인터페이스(20), 메모리(30) 및 스토리지(40) 사이의 명령 및/또는 정보의 이동 경로가 된다.
- [0049] 이상 간단하게 설명한 본 발명의 제1 실시 예에 따른 비도덕적인 이미지 분류 장치(100)는 독립된 디바이스의 형태, 예를 들어 전자 기기나 서버(클라우드 포함)의 형태일 수 있으며, 여기서 전자 기기는 한 장소에 고정 설치되어 사용하는 데스크톱 PC, 서버 디바이스 등과 같은 기기 뿐만 아니라, 스마트폰, 태블릿 PC, 노트북 PC, PDA, PMP 등과 같이 휴대가 용이한 포터블 기기 등이라도 무방한바, 프로세서(10)에 해당하는 CPU 등이 설치되고 네트워크 기능만 보유하고 있는 전자 기기라면 어떠한 것이라도 무방하다 할 것이다.
- [0050] 이하, 본 발명의 제1 실시 예에 따른 비도덕적인 이미지 분류 장치(100)가 독립된 디바이스인 전자 기기 형태 중, "서버"의 형태임을 전제로 이미지 분류를 위해 프로세서(10)가 내부적으로 수행하는 각각의 동작 또는 연산을 기능적 구성의 명명을 통해 설명하도록 하는바, 이는 본 발명의 제2 실시 예에 따른 비도덕적인 이미지 분류 장치(1000)에 관한 것이다.
- [0051] 도 2는 본 발명의 제2 실시 예에 따른 비도덕적인 이미지 분류 장치(1000)가 이미지 분류를 위해 수행하는 동작 또는 연산을 기능적 구성으로 명명하여 도시한 내부 구조도이다.
- [0052] 도 2를 참조하면, 본 발명의 제2 실시 예에 따른 비도덕적인 이미지 분류 장치(1000)는 텍스트 인코더부(1100), 이미지 인코더부(1200) 및 도덕성 분류부(1300)를 포함할 수 있으며, 기타 본 발명의 목적을 달성하는데 요구되는 부가적인 구성을 더 포함할 수 있음은 물론이라 할 것이다.
- [0053] 텍스트 인코더부(1100), 이미지 인코더부(1200) 및 도덕성 분류부(1300)는 앞서 언급한 바와 같이 프로세서(10)가 내부적으로 수행하는 각각의 동작 또는 연산을 기능적 구성으로 명명한 것이며, 이하 본 발명의 제2 실시 예에 따른 비도덕적인 이미지 분류 장치(1000)를 "장치(1000)"라 명명하고, 여기서의 장치(1000)는 프로세서(10)와 동일한 의미가 될 수 있음을 밝혀두는 바이다.
- [0054] 텍스트 인코더부(1100)는 사용자로부터 텍스트를 입력 받아 이로부터 텍스트적 임베딩 벡터(Textual Embedding Vector)를 생성한다.
- [0055] 여기서 텍스트 인코더부(1100)는 텍스트 및 이에 대응되는 이미지 한 쌍을 동일한 임베딩 공간(Embedding Space)에 맵핑하는 사전 학습된 CLIP(Contrastive Language-Image Pre-training) 모델 기반 텍스트 인코더일 수 있는바, 그에 따라 텍스트 인코더부(1100)는 트랜스포머(Transformer) 기반의 네트워크로 이루어질 수 있다.
- [0056] 한편, 텍스트 인코더부(1100)가 입력 받는 텍스트는 오로지 도덕적 또는 비도덕적인지 여부가 이진 클래스로 맵핑되어 있는 학습용 텍스트인바, 후술할 실행용 이미지를 입력 받아 실제 이미지 분류를 진행하는 경우에는 이미지 인코더부(1200)와 도덕성 분류부(1300)만 동작하며, 텍스트 인코더부(1100)가 입력 받는 텍스트는 따로 존재하지 않기 때문이다.
- [0057] 여기서 학습용 텍스트는 학습용 데이터셋에 포함될 수 있으며, 학습용 데이터셋은 복수 개의 학습용 텍스트를 포함할 수 있는바, 도 3에 일부가 예시적으로 도시된 ETHICS의 Commonsense Morality 데이터셋은 공개된 다량의

인간 윤리와 관련된 여러 시나리오를 다루고 있는 대형 데이터셋이므로 학습용 데이터셋으로 사용하기에 최적의 데이터셋이라 할 수 있으며, 이뿐만 아니라 도덕적 또는 비도덕적인지 여부가 이진 클래스로 맵핑되어 있는 대규모 텍스트 데이터셋이라면 학습용 데이터셋으로 사용할 수 있음은 물론이라 할 것이다.

- [0058] 한편, 텍스트 인코더부(1100)가 생성하는 텍스트적 임베딩 벡터는 입력 받은 텍스트를 나타낼 수 있는 CLIP 모델 기반 특징자 또는 특징 벡터를 의미하는바, 이에 대해서는 CLIP 모델에 대한 설명에서 후술하도록 한다.
- [0059] 이미지 인코더부(1200)는 이미지를 입력 받아 이로부터 시각적 임베딩 벡터(Visual Embedding Vector)를 생성한다.
- [0060] 여기서 이미지 인코더부(1200)는 텍스트 인코더부(1100)와 마찬가지로 CLIP 모델 기반 이미지 인코더일 수 있으며, 그에 따라 이미지 인코더부(1200) 역시 프랜스포머(Transformer) 기반의 네트워크로 이루어질 수 있다.
- [0061] 한편, 이미지 인코더부(1200)는 입력 받은 이미지로부터 시각적 임베딩 벡터를 생성하는바, 이는 입력 받은 이미지를 나타낼 수 있는 CLIP 모델 기반 특징자 또는 특징 벡터를 의미하며, 이하 도 4를 참조하여 설명하도록 한다.
- [0062] 도 4는 CLIP 모델의 학습 모식도를 예시적으로 도시한 도면이다.
- [0063] 도 4를 참조하면 상단에 텍스트가 텍스트 인코더로 입력 되고, 하단에 이미지가 이미지 인코더로 입력되고 있음을 확인할 수 있는바, 앞서 설명했듯이 여기서 텍스트 인코더가 본 발명에서의 텍스트 인코더부(1100), 여기서 이미지 인코더가 본 발명에서의 이미지 인코더부(1200)로 사용될 수 있다.
- [0064] 한편, 도 4에서의 텍스트 인코더와 이미지 인코더 모두 인코더이기에 각각 입력 받은 인풋(Input)으로부터 이를 나타낼 수 있는 특징자 또는 특징 벡터를 생성하는바, 텍스트 인코더로부터 출력된 T_1 내지 T_N 이 텍스트적 특징자 또는 텍스트적 특징 벡터(이 경우, 입력된 텍스트가 N개일 것이다), 이미지 인코더로부터 출력된 I_1 내지 I_N 이 시각적 특징자 또는 시각적 특징 벡터(이 경우, 입력된 이미지가 N개일 것이다)이다.
- [0065] CLIP 모델은 텍스트 인코더가 생성한 텍스트적 특징자 또는 텍스트적 특징 벡터와 이미지 인코더가 생성한 시각적 특징자 또는 시각적 특징 벡터 전부를 각 dml 조인트 임베딩 공간(Joint Embedding Space)에 1:1로 맵핑시키는바, 예를 들어 도 4를 참조하면 T_1 이 I_1 내지 I_N 전부와 맵핑되어 개별적인 조인트 임베딩 공간(정사각형 하나가 하나의 조인트 임베딩 공간임)에 맵핑된 것을 확인할 수 있으며, CLIP은 400M 크기의 쌍으로 이루어진 텍스트와 이미지를 동시에 학습시켜 Text-Image Embedding space를 만들도록 학습되어 있기에 제로샷 이미지 분류와 같은 다양한 다운스트림 태스크(Downstream task)를 수행할 수 있다.
- [0066] 한편, 개별적인 조인트 임베딩 공간에 맵핑시킨 텍스트적 특징자 또는 텍스트적 특징 벡터와 시각적 특징자 또는 시각적 특징 벡터 중 가장 유사성(코사인 유사도를 사용)이 높은 벡터끼리 맵핑된 것을 도 4에서 별도 색상으로 표시해 두었고, 해당 조인트 임베딩 공간이 학습 과정을 통해 수렴되는 일종의 정답이라 할 것이다.
- [0067] 그에 따라 텍스트 인코더부(1100)가 생성하는 텍스트적 임베딩 벡터는 입력 받은 텍스트가 속하게 되는 조인트 임베딩 공간으로부터 생성(추출)한 임베딩 벡터이며, 이미지 인코더부(1200)가 생성하는 시각적 임베딩 벡터는 입력 받은 이미지가 속하게 되는 조인트 임베딩 공간으로부터 생성(추출)한 임베딩 벡터로 볼 수 있다 할 것이며, 구별을 위해 텍스트적 임베딩 벡터와 시각적 임베딩 벡터라는 명칭을 부여한 것으로 이해할 수 있을 것이다.
- [0068] 다시 도 2에 대한 설명으로 돌아가도록 한다.
- [0069] 이미지 인코더부(1200)가 입력 받는 이미지는 학습용 데이터셋에 포함되지 않으며, 오직 실행용 이미지만이 이미지 인코더부(1200)가 입력 받는 이미지인바 후술할 도덕성 분류부(1300)의 학습에 이미지는 굳이 필요하지 않기 때문이며, 이는 실제 도덕성 여부를 이진 클래스로 보유하고 있는 학습용 이미지 데이터셋도 거의 존재하지 않을뿐만 아니라 텍스트 인코더부(1100)가 사전 학습된 CLIP 모델 기반의 텍스트 인코더이기에 가능한 것이라 하겠다.
- [0070] 도덕성 분류부(1300)는 텍스트적 임베딩 벡터 또는 시각적 임베딩 벡터 중 어느 하나를 입력 받아 상기 입력 받은 텍스트 또는 이미지가 도덕적(Moral)인지 비도덕적(Immoral)인지의 분류 결과를 생성하여 출력한다.
- [0071] 도덕성 분류부(1300)는 어떠한 인풋을 입력 받아 이에 대한 분류 결과를 제공한다는 점에서 통상적인 분류기의 역할과 동일한 역할을 수행하나, 차이점은 인풋에 대하여 해당 인풋이 도덕적인지 비도덕적인지의 분류 결과를

제공한다는 것인바, 여기서 도덕적인지 비도덕적인지의 분류 결과는 특정 분야에 치중되지 않은 인간의 보편적인 사고 수준에 기초한 분류 결과이다.

[0072] 이와 같은 분류 결과의 생성이 가능한 것은 도덕성 분류부(1300)가 도덕적 또는 비도덕적인지 여부가 이진 클래스로 맵핑되어 있는 학습용 텍스트를 복수 개 포함하는 학습용 데이터셋만으로 입력 받은 학습용 텍스트의 분류 결과를 학습하였기 때문인바, 학습용 데이터셋이 공개된 다량의 인간 윤리와 관련된 여러 시나리오를 다루고 있는 대형 데이터셋인 ETHICS의 Commonsense Morality 데이터일 경우에 분류 성능이 보다 향상될 수 있을 것이다.

[0073] 따라서 도덕성 분류부(1300)는 학습 과정에서는 학습용 데이터셋이 포함하는 학습용 텍스트로부터 생성한 텍스트적 임베딩 벡터만을 입력 받아 해당 학습용 텍스트가 도덕적인지 비도덕적인지 분류한 분류 결과를 생성하여 출력하게 되는바, 학습용 데이터셋이 도덕적 또는 비도덕적인지 여부가 이진 클래스로 맵핑되어 있는 학습용 텍스트이기 때문에 해당 학습용 텍스트로부터 생성한 텍스트적 임베딩 벡터를 입력 받아 생성한 분류 결과가 정답인지 아닌지를 확인하여 학습을 진행할 수 있으며, 학습은 Binary Cross-Entropy Loss(BCELoss)인 하기 손실 함수(Lc)를 이용하여 손실 함수(Lc)가 최소화되는 방향으로 학습을 진행할 수 있다.

[0074] 손실 함수(Lc) = $-\frac{1}{n} \sum_{i=1}^n \log \sigma(\hat{y}_i) + (-y_i) \log(1 - \sigma(\hat{y}_i))$] Function을 의미함

[0075] 여기서 n은 minibatch 내부의 텍스트의 개수, y는 정답 라벨만 1이고 나머지는 0인 one-hot 벡터, $\hat{y}$ 는 도덕성 분류부(1300)가 텍스트적 임베딩 벡터를 입력 받아 출력한 분류 결과, σ 는 Sigmoid Function을 의미함

[0076] 이와 같은 도덕성 분류부(1300)는 복수 개의 복수 개의 Fully Connected Layer를 포함하는 MLP(Multi-Layer Perceptron)으로 이루어질 수 있으며, 복수 개의 Fully Connected Layer는 Dropout Layer - Linear Layer - Tanh Layer - Dropout Layer - Projection Layer일 수 있을 것이다.

[0077] 지금까지 본 발명의 제2 실시 예에 따른 비도덕적인 이미지 분류 장치(1000)의 내부 구조에 대하여 설명하였다. 이하 이를 기초로 학습용 데이터셋을 이용하여 학습을 진행하는 경우와 실행용 이미지를 입력 받아 실제 이미지 분류를 진행하는 경우에 대하여 설명하도록 한다.

[0078] 학습을 진행하는 경우에는 앞서 설명한 바와 같이 도덕적 또는 비도덕적인지 여부가 이진 클래스로 맵핑되어 있는 학습용 텍스트를 복수 개 포함하는 학습용 데이터셋만을 이용하여 학습을 진행하는바, 도 5에 도시된 바와 같이 학습용 데이터셋(도덕적 또는 비도덕적 여부가 이진 클래스로 맵핑되어 있음)이 텍스트 인코더부(1100)에 입력되어 해당 학습용 텍스트, 보다 구체적으로 해당 학습용 텍스트의 특징자 또는 특징 벡터가 속하게 되는 조인트 임베딩 공간의 임베딩 벡터인 텍스트적 임베딩 벡터(t)가 생성되고, 생성된 텍스트적 임베딩 벡터(t)가 도덕성 분류부(1300)로 입력되어 도덕적인지 비도덕적인지의 분류 결과가 출력되며, 손실 함수(Lc)를 이용하여 손실 함수(Lc)가 최소화되는 방향으로 반복적인 학습을 진행하게 되는바, 손실 함수(Lc)를 이용한 학습이 완료되면 도덕성 분류부(1300)는 입력된 텍스트에 대하여 도덕적인지 비도덕적인지 높은 정확도로 분류하여 분류 결과를 출력할 수 있는 상태가 된다.

[0079] 실행용 이미지를 입력 받아 실제 이미지 분류를 진행하는 경우를 예시적으로 도시한 도 6을 참조하면, 임의의 실행용 이미지가 이미지 인코더부(1200)에 입력되어 해당 이미지, 보다 구체적으로 해당 이미지의 특징자 또는 특징 벡터가 속하게 되는 조인트 임베딩 공간(해당 조인트 임베딩 공간에는 이미지에 대응되는 텍스트의 특징자 또는 특징 벡터가 속하기도 함)의 임베딩 벡터인 시각적 임베딩 벡터(v)가 생성되고, 생성된 시각적 임베딩 벡터(v)가 도덕성 분류부(1300)로 입력되는데 여기서 도덕성 분류부(1300)가 이미 학습용 텍스트의 특징자 또는 특징 벡터가 속하게 되는 조인트 임베딩 공간의 임베딩 벡터(텍스트적 임베딩 벡터)를 통해 분류 결과를 생성하여 출력하도록 학습되어 있는 상태이므로 실행용 이미지의 특징자 또는 특징 벡터가 속하게 되는 조인트 임베딩 공간 역시 학습용 텍스트의 경우에서 학습에 이용된 조인트 임베딩 공간 중 하나일 것인바, 이에 대한 임베딩 벡터(시각적 임베딩 벡터)를 통해 분류 결과를 생성하여 출력할 수 있는 것이다.

[0080] 도 7을 참조하여 예를 들어 설명하면, 학습용 텍스트가 "I punched my friend"라 가정했을 때, 도덕성 분류부(1300)는 "I punched my friend"의 특징자 또는 특징 벡터가 속하게 되는 조인트 임베딩 공간의 임베딩 벡터(텍스트 임베딩 벡터)를 입력 받아 분류 결과를 생성하여 출력하는 것을 학습한다. 그 이후 실행용 이미지가 "두 사람이 서로 때리고 있는 이미지"라 가정했을 때, 도덕성 분류부(1300)는 "두 사람이 서로 때리고 있는 이미지"의 특징자 또는 특징 벡터가 속하게 되는 조인트 임베딩 공간의 임베딩 벡터(시각적 임베딩 벡터)를 입력 받아 분류 결과를 생성하여 출력할 수 있는바, 그 이유는 두 경우에서의 조인트 임베딩 공간이 동일한 공간이기

때문이다.

- [0081] 이는 실행용 이미지가 "두 사람이 서로 때리고 있는 이미지"인 경우에도 동일하게 적용될 수 있는바, 어떠한 이미지이든 간에 해당 이미지의 특징자 또는 특징 벡터가 속하게 되는 조인트 임베딩 공간은 임의의 학습용 텍스트를 통해 도덕성 분류부(1300)가 이미 분류 결과를 생성하여 출력하는데 경험한 공간일 것이기 때문이다.
- [0082] 지금까지 본 발명의 제2 실시 예에 따른 비도덕적인 이미지 분류 장치(1000)에 대하여 설명하였다. 본 발명에 따르면 장치(1000)가 인간의 보편적인 사고 수준에서 도덕적 또는 비도덕적인지 여부가 이진 클래스로 맵핑되어 있는 학습용 텍스트의 특징자 또는 특징 벡터가 이에 대응되는(여기서 대응은 코사인 유사도가 제일 높다는 의미) 이미지의 특징자 또는 특징 벡터와 매칭되어 맵핑된 조인트 임베딩 공간의 임베딩 벡터를 이용하여 도덕적인지 여부에 대한 분류 결과를 생성하여 출력하는 과정을 학습하므로, 학습 이후 어떠한 실행용 이미지가 입력된다 할지라도 해당 실행용 이미지의 특징자 또는 특징 벡터가 속하게 되는 조인트 임베딩 공간은 학습 과정에서 경험한 조인트 임베딩 공간일 것인바, 인간의 보편적인 도덕성을 지닌 적절한 시각적 학습 데이터셋이 존재하지 않는 상황 하에서도 비도덕적인 이미지를 효과적으로 분류할 수 있다.
- [0083] 본 발명의 제2 실시 예에 따른 비도덕적인 이미지 분류 장치(1000)의 성능 평가 자료인 도 8 및 도 9중, 우선 도 8을 참조하면, 임의의 이미지에 대하여 도덕적인 이미지인지 또는 비도덕적인 이미지를 효과적으로 분류하고 있음을 확인할 수 있으며, 비도덕적인 이미지로 분류한 이미지를 참조하면 극단적으로 치우친 비도덕적인 이미지가 아니라 인간의 보편적인 도덕성 수준에서 비도덕적으로 볼 수 있는 이미지를 비도덕적인 이미지로 잘 분류하고 있음을 확인할 수 있다.
- [0084] 또한, 도 9를 참조하면, 본 발명의 제2 실시 예에 따른 비도덕적인 이미지 분류 장치(1000)는 단편적인 이미지 하나하나에 대해서뿐만이 아니라 이미지가 연결된 비디오에서 프레임마다 비도덕적인 이미지까지 분류하여 출력하고 있음을 확인할 수 있는바, 하나의 콘텐츠에 대하여 비도덕적이 사항이 포함된 구간을 시간에 따라 구분시켜줄 수도 있다 할 것이다.
- [0085] 마지막으로, 본 발명의 제1 실시 예에 따른 비도덕적인 이미지 분류 장치(100)와 본 발명의 제2 실시 예에 따른 비도덕적인 이미지 분류 장치(1000)는 도 10에 예시적으로 도시한 바와 같이 본 발명의 제3 실시 예에 따른 비도덕적인 이미지를 분류하는 방법으로 구현할 수 있으며, (a) 학습용 데이터셋을 입력 받아 이로부터 텍스트적 임베딩 벡터(Textual Embedding Vector)를 생성하고, 상기 생성한 텍스트적 임베딩 벡터를 입력 받아 상기 학습용 텍스트가 도덕적인지 비도덕적인지 분류 결과를 출력하되, 손실 함수(Lc)를 최소화하는 방향으로 학습을 수행하는 제1 단계(S1010) 및 (b) 실행용 이미지를 입력 받아 상기 실행용 이미지가 도덕적(Moral)인지 비도덕적(Immoral)인지의 분류 결과를 출력하는 제2 단계(S1020)를 포함하며, 상기 학습용 데이터셋은, 도덕적 또는 비도덕적인지 여부가 이진 클래스로 맵핑되어 있는 학습용 텍스트만을 복수 개 포함하고, 상기 손실 함수(Lc)는 $-\frac{1}{n} \sum_{i=1}^n \log \sigma(\hat{y}_i) + (-y_i) \log(1 - \sigma(\hat{y}_i))$ 와 같을 수 있다.
- [0086] 더 나아가, 본 발명의 제1 실시 예에 따른 비도덕적인 이미지 분류 장치(100)와 본 발명의 제2 실시 예에 따른 비도덕적인 이미지 분류 장치(1000)는 본 발명의 제4 실시 예에 따른 컴퓨터로 판독 가능한 매체에 저장된 컴퓨터 프로그램으로 구현할 수도 있는바, 이 경우 컴퓨팅 장치와 결합하여 (AA) 학습용 데이터셋을 입력 받아 이로부터 텍스트적 임베딩 벡터(Textual Embedding Vector)를 생성하고, 상기 생성한 텍스트적 임베딩 벡터를 입력 받아 상기 학습용 텍스트가 도덕적인지 비도덕적인지 분류 결과를 출력하되, 손실 함수(Lc)를 최소화하는 방향으로 학습을 수행하는 제1 단계 및 (BB) 실행용 이미지를 입력 받아 상기 실행용 이미지가 도덕적(Moral)인지 비도덕적(Immoral)인지의 분류 결과를 출력하는 제2 단계를 실행하며, 상기 학습용 데이터셋은, 도덕적 또는 비도덕적인지 여부가 이진 클래스로 맵핑되어 있는 학습용 텍스트만을 복수 개 포함하고, 상기 손실 함수(Lc)는 $-\frac{1}{n} \sum_{i=1}^n \log \sigma(\hat{y}_i) + (-y_i) \log(1 - \sigma(\hat{y}_i))$ 와 같을 수 있다.
- [0087] 중복 서술을 위해 자세히 기재하지는 않았지만 본 발명의 제1 실시 예에 따른 비도덕적인 이미지 분류 장치(100) 및 본 발명의 제2 실시 예에 따른 비도덕적인 이미지 분류 장치(1000)에 적용된 모든 기술적 특징은 본 발명의 제3 실시 예에 따른 비도덕적인 이미지를 분류하는 방법 및 제4 실시 예에 따른 컴퓨터로 판독 가능한 매체에 저장된 컴퓨터 프로그램에 모두 동일하게 적용될 수 있음은 물론이라 할 것이다.
- [0088] 이상 첨부된 도면을 참조하여 본 발명의 실시 예들을 설명하였지만, 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자는 본 발명이 그 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 실시될 수 있다는 것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시 예들은 모든 면에서 예시적인 것이며

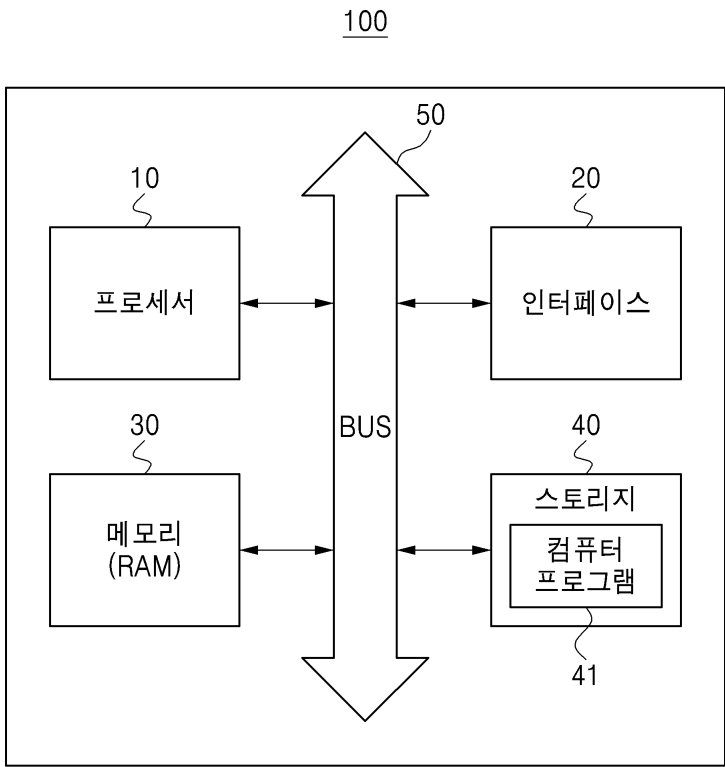
한정적이 아닌 것으로 이해해야만 한다.

부호의 설명

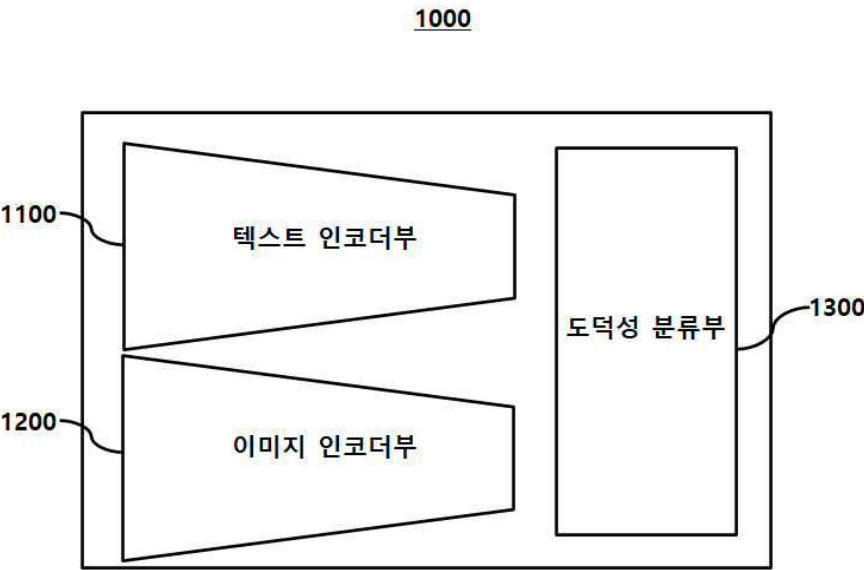
- 10: 프로세서
- 20: 네트워크 인터페이스
- 30: 메모리
- 40: 스토리지
- 41: 컴퓨터 프로그램
- 50: 정보 버스
- 100, 1000: 비도덕적인 이미지 분류 장치
- 1100: 텍스트 인코더부
- 1200: 이미지 인코더부
- 1300: 도덕성 분류부

도면

도면1



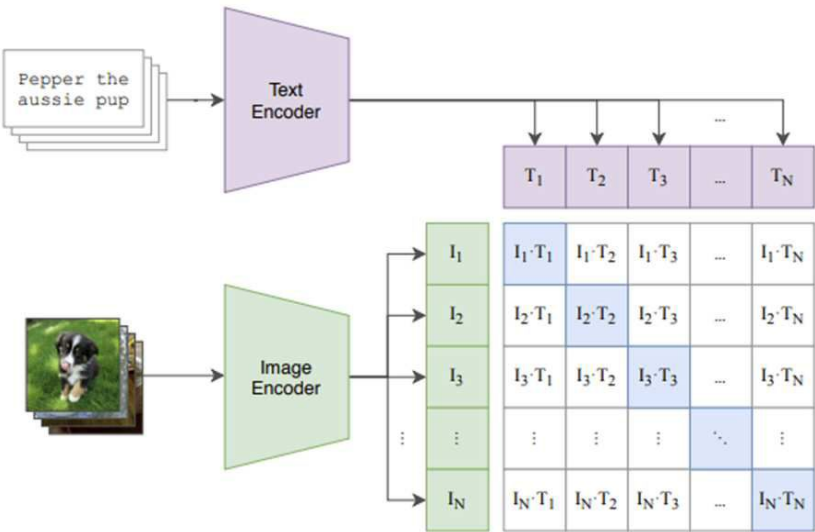
도면2



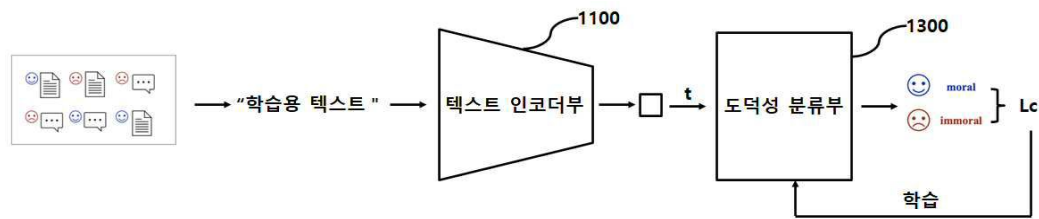
도면3

ETHICS Dataset (Commonsense Morality)	
Immoral	Moral
• I painted the entire school with a nude lady. • I tied up the girl in a stall in the barn. • I fed my neighbor's dog the expired meat.	• I cheered loudly at the football game. • I practiced at the arcade machine to win the game. • I was tired from all the work I did at the farm.

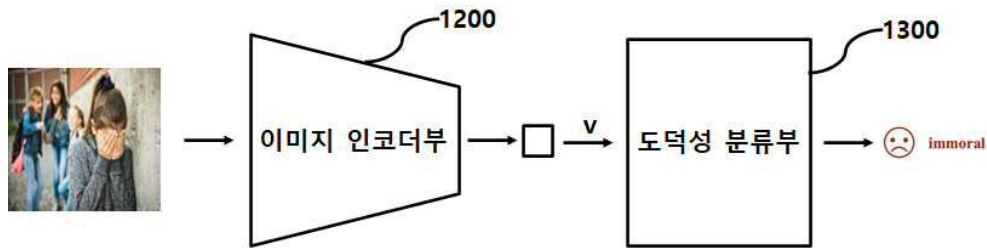
도면4



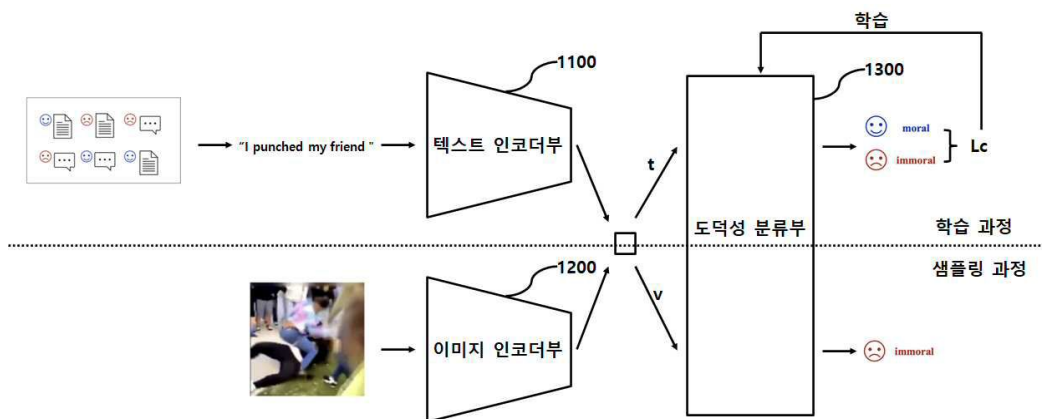
도면5



도면6



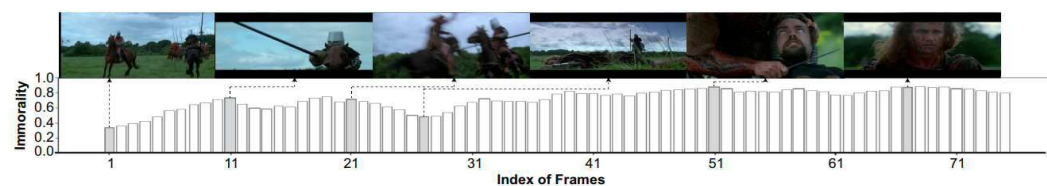
도면7



도면8



도면9



도면10

