(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2023년10월24일
(11) 등록번호 10-2593136
(24) 등록일자 2023년10월19일

(51) 국제특허분류(Int. Cl.)
G06T 3/40 (2006.01) *G06F 16/53* (2019.01)
G06F 16/583 (2019.01) *G06T 11/00* (2006.01)
G06T 5/00 (2019.01)
(52) CPC특허분류
G06T 3/4038 (2013.01)
G06F 16/53 (2019.01)
(21) 출원번호 10-2023-0047710
(22) 출원일자 2023년04월11일
심사청구일자 2023년04월11일
(56) 선행기술조사문헌
KR102358186 B1*
Lvmin Zhang et al., "Adding Conditional Control to Text-to-Image Diffusion Models", Computer Vision and Pattern Recognition, arXiv:2302.05543v1 [cs.CV], (2023.02.10.)*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
고려대학교산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
김진규
서울시 광진구 아차산로 599 신동아파밀리에 107동 801호
박성범
경기도 화성시봉담읍 동화새터길 55-39 (동화마을 신동아파밀리에아파트) 108동 1301호
(74) 대리인
특허법인위솔

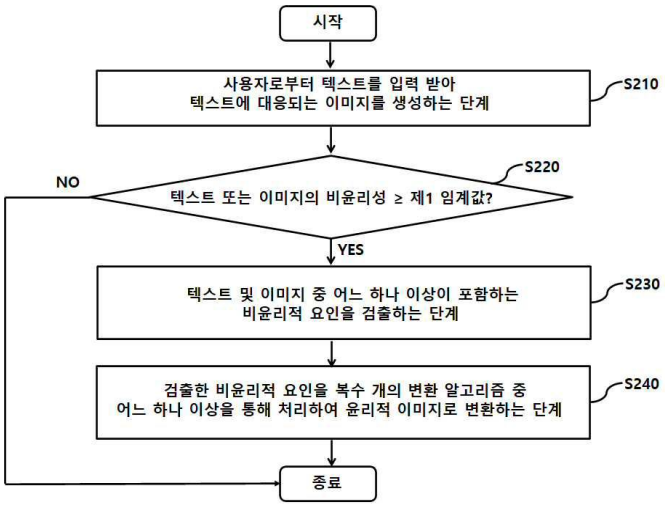
전체 청구항 수 : 총 8 항 심사관 : 이정은

(54) 발명의 명칭 **비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 방법 및 이를 위한 장치**

(57) 요 약

본 발명의 일 실시 예에 따른 비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 방법은 (a) 사용자로부터 텍스트를 입력 받아 상기 텍스트에 대응되는 이미지를 생성하는 제1 단계, (b) 상기 입력 받은 텍스트 및 상기 생성한 이미지 중 어느 하나 이상의 비윤리성을 측정하여, 상기 측정한 비윤리성이 제1 임계값 이상인지 판단하는 제2 단계, (c) 상기 판단 결과 비윤리성이 제1 임계값 이상이라면, 상기 텍스트 및 이미지 중 어느 하나 이상이 포함하는 비윤리적 요인을 검출하는 제3 단계 및 (d) 상기 검출한 비윤리적 요인을 복수 개의 변환 알고리즘(Algorithm) 중 어느 하나 이상을 통해 처리하여 윤리적 이미지로 변환하는 제4 단계를 포함하며, 상기 복수 개의 변환 알고리즘은, 모자이크(Mosaic) 처리 알고리즘, 인페인팅(Inpainting) 처리 알고리즘, 대체 단어 생성 알고리즘 및 윤리적 캡션(Caption) 활용 알고리즘 중 어느 하나 이상이다.

대 표 도 - 도2



(52) CPC특허분류
G06F 16/5846 (2019.01)
G06T 11/00 (2013.01)
G06T 5/005 (2023.01)

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	1/1
과제수행기관명	고려대학교산학협력단
연구기간	2023.01.01 ~ 2023.12.31

공지예외적용 : 있음

명 세 서

청구범위

청구항 1

프로세서 및 메모리를 포함하는 장치가 비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 방법에 있어서,

(a) 사용자로부터 텍스트를 입력 받아 상기 텍스트에 대응되는 이미지를 생성하는 제1 단계;

(b) 상기 입력 받은 텍스트 및 상기 생성한 이미지 중 어느 하나 이상의 비윤리성을 윤리 관련 텍스트 데이터셋인 ETHICS 데이터 셋 및 이미지-텍스트 임베딩 모델인 CLIP 모델을 이용하여 학습이 완료된 비윤리성 분류기(Immorallity Classifier)를 이용하여 측정하여, 상기 측정한 비윤리성이 제1 임계값 이상인지 판단하는 제2 단계;

(c) 상기 판단 결과 비윤리성이 제1 임계값 이상이라면, 상기 텍스트 및 이미지 중 어느 하나 이상이 포함하는 비윤리적 요인을 검출하는 제3 단계; 및

(d) 상기 검출한 비윤리적 요인을 복수 개의 변환 알고리즘(Algorithm) 중 어느 하나 이상을 통해 처리하여 윤리적 이미지로 변환하는 제4 단계;

를 포함하며,

상기 복수 개의 변환 알고리즘은,

모자이크(Mosaic) 처리 알고리즘, 인페인팅(Inpainting) 처리 알고리즘, 대체 단어 생성 알고리즘 및 윤리적 캡션(Caption) 활용 알고리즘 중 어느 하나 이상인,

비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 방법에 있어서,

상기 제3 단계는,

상기 이미지가 포함하는 픽셀 중, 복수 개의 픽셀을 랜덤으로 마스킹한 하나 이상의 마스킹 이미지(Ⅱ)를 생성하는 제3-1´ 단계;

상기 생성한 하나 이상의 마스킹 이미지(Ⅱ)의 비윤리성을 측정하는 제3-2´ 단계;

상기 제3-2´ 단계에서 측정한 마스킹 이미지(Ⅱ)의 비윤리성을 이용하여 상기 이미지가 포함하는 각각의 픽셀별 비윤리성을 산정하는 제3-3´ 단계; 및

상기 산정한 픽셀 별 비윤리성이 제2 임계값 이상인 픽셀을 비윤리적 요인으로 검출하는 제3-4 ´단계;

중 어느 하나 이상을 포함하는 비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 방법.

청구항 2

제1항에 있어서,

상기 제1 단계에서 이미지의 생성은,

조건부 확산(Conditional Latent Diffusion) 모델에 기반한 스테이블 디퓨전 모델(Stable Diffusion)을 이용하여 생성하는,

비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 방법.

청구항 3

삭제

청구항 4

제1항에 있어서,

상기 제3 단계는,

상기 텍스트를 단어 단위로 분할하는 제3-1 단계;

상기 단어 단위로 분할한 텍스트가 포함하는 단어 중 어느 하나를 마스킹하여 하나 이상의 마스킹 텍스트를 생성하는 제3-2 단계;

상기 생성한 하나 이상의 마스킹 텍스트에 대응되는 하나 이상의 이미지(Ⅰ)를 생성하는 제3-3 단계;

상기 생성한 마스킹 텍스트에 대응되는 하나 이상의 이미지(Ⅰ)의 비윤리성을 측정하는 제3-4 단계; 및

상기 제3-4 단계에서 측정한 이미지(Ⅰ)의 비윤리성과 상기 제2 단계에서 측정한 이미지의 비윤리성을 비교하여 비윤리적 요인을 검출하는 제3-5 단계;

중 어느 하나 이상을 포함하는 비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 방법.

청구항 5

제4항에 있어서,

상기 제3-5 단계는,

상기 제3-4 단계에서 측정한 이미지(Ⅰ)의 비윤리성과 상기 제2 단계에서 측정한 이미지의 비윤리성의 차이의 절대값이 가장 큰 이미지(Ⅰ)를 생성하는데 이용한 마스킹 텍스트에서 마스킹된 단어를 비윤리적 요인으로 검출하는,

비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 방법.

청구항 6

삭제

청구항 7

제 1항에 있어서,

상기 제3-3´ 단계는,

상기 마스킹 이미지(Ⅱ)가 포함하는 마스킹된 복수 개의 픽셀 중 어느 하나 및 또 다른 마스킹 이미지(Ⅱ)가 포함하는 상기 픽셀과 동일한 위치의 마스킹되지 않은 픽셀을 선택하되, 상기 마스킹 이미지(Ⅱ) 및 또 다른 마스킹 이미지(Ⅱ)에 대하여 상기 제3-2´단계에서 측정한 비윤리성을 가중 합계(Weighted SUM)하여 상기 선택한 마스킹된 복수 개의 픽셀 중 어느 하나에 대한 비윤리성을 산정하는,

비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 방법.

청구항 8

제1항에 있어서,

상기 제3-4´ 단계 이후에,

상기 검출한 비윤리적 요인을 포함하는 상기 이미지를 히트맵 이미지로 변환하는 제3-5´ 단계;

를 더 포함하는 비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 방법.

청구항 9

하나 이상의 프로세서;

네트워크 인터페이스;

상기 프로세서에 의해 수행되는 컴퓨터 프로그램을 로드(Load)하는 메모리; 및

대용량 네트워크 데이터 및 상기 컴퓨터 프로그램을 저장하는 스토리지를 포함하되,

상기 컴퓨터 프로그램은 상기 하나 이상의 프로세서에 의해,

(A) 사용자로부터 텍스트를 입력 받아 상기 텍스트에 대응되는 이미지를 생성하는 제1 오퍼레이션;

(B) 상기 입력 받은 텍스트 및 상기 생성한 이미지 중 어느 하나 이상의 비윤리성을 윤리 관련 텍스트 데이터셋인 ETHICS 데이터 셋 및 이미지-텍스트 임베딩 모델인 CLIP 모델을 이용하여 학습이 완료된 비윤리성 분류기(Immorallity Classifier)를 이용하여 측정하여, 상기 측정한 비윤리성이 제1 임계값 이상인지 판단하는 제2 오퍼레이션;

(C) 상기 판단 결과 비윤리성이 제1 임계값 이상이라면, 상기 텍스트 및 이미지 중 어느 하나 이상이 포함하는 비윤리적 요인을 검출하는 제3 오퍼레이션; 및

(D) 상기 검출한 비윤리적 요인을 복수 개의 변환 알고리즘(Algorithm) 중 어느 하나 이상을 통해 처리하여 윤리적 이미지로 변환하는 제4 오퍼레이션;

을 실행하며,

상기 복수 개의 변환 알고리즘은,

모자이크(Mosaic) 처리 알고리즘, 인페인팅(Inpainting) 처리 알고리즘, 대체 단어 생성 알고리즘 및 윤리적 캡션(Caption) 활용 알고리즘 중 어느 하나 이상인,

비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 장치에 있어서,

상기 제3 오퍼레이션은,

상기 이미지가 포함하는 픽셀 중, 복수 개의 픽셀을 랜덤으로 마스킹한 하나 이상의 마스킹 이미지(Ⅱ)를 생성하는 제3-1´ 오퍼레이션;

상기 생성한 하나 이상의 마스킹 이미지(Ⅱ)의 비윤리성을 측정하는 제3-2´ 오퍼레이션;

상기 제3-2´ 오퍼레이션에서 측정한 마스킹 이미지(Ⅱ)의 비윤리성을 이용하여 상기 이미지가 포함하는 각각의 픽셀 별 비윤리성을 산정하는 제3-3´ 오퍼레이션; 및

상기 산정한 픽셀 별 비윤리성이 제2 임계값 이상인 픽셀을 비윤리적 요인으로 검출하는 제3-4 ´ 오퍼레이션;

중 어느 하나 이상을 포함하는 비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 장치.

청구항 10

컴퓨팅 장치와 결합하여,

(AA) 사용자로부터 텍스트를 입력 받아 상기 텍스트에 대응되는 이미지를 생성하는 제1 단계;

(BB) 상기 입력 받은 텍스트 및 상기 생성한 이미지 중 어느 하나 이상의 비윤리성을 윤리 관련 텍스트 데이터셋인 ETHICS 데이터 셋 및 이미지-텍스트 임베딩 모델인 CLIP 모델을 이용하여 학습이 완료된 비윤리성 분류기(Immorallity Classifier)를 이용하여 측정하여, 상기 측정한 비윤리성이 제1 임계값 이상인지 판단하는 제2 단계;

(CC) 상기 판단 결과 비윤리성이 제1 임계값 이상이라면, 상기 텍스트 및 이미지 중 어느 하나 이상이 포함하는 비윤리적 요인을 검출하는 제3 단계; 및

(DD) 상기 검출한 비윤리적 요인을 복수 개의 변환 알고리즘(Algorithm) 중 어느 하나 이상을 통해 처리하여 윤리적 이미지로 변환하는 제4 단계;

를 포함하며,

상기 복수 개의 변환 알고리즘은,

모자이크(Mosaic) 처리 알고리즘, 인페인팅(Inpainting) 처리 알고리즘, 대체 단어 생성 알고리즘 및 윤리적 캡션(Caption) 활용 알고리즘 중 어느 하나 이상인,

컴퓨터로 판독 가능한 매체에 저장된 컴퓨터 프로그램에 있어서,

상기 제3 단계는,

상기 이미지가 포함하는 픽셀 중, 복수 개의 픽셀을 랜덤으로 마스킹한 하나 이상의 마스킹 이미지(Ⅱ)를 생성

하는 제3-1´ 단계;

상기 생성한 하나 이상의 마스킹 이미지(Ⅱ)의 비윤리성을 측정하는 제3-2´ 단계;

상기 제3-2´ 단계에서 측정한 마스킹 이미지(Ⅱ)의 비윤리성을 이용하여 상기 이미지가 포함하는 각각의 픽셀 별 비윤리성을 산정하는 제3-3´ 단계; 및

상기 산정한 픽셀 별 비윤리성이 제2 임계값 이상인 픽셀을 비윤리적 요인으로 검출하는 제3-4 ´단계;

중 어느 하나 이상을 포함하는 컴퓨터로 판독 가능한 매체에 저장된 컴퓨터 프로그램.

발명의 설명

기 술 분 야

[0001] 본 발명은 비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 방법 및 이를 위한 장치에 관한 것이다. 보다 자세하게는 이미지의 비윤리성을 파악하여 비윤리적인 이미지를 윤리적인 이미졸 변환하는 방법 및 이를 위한 장치에 관한 것이다.

배 경 기 술

[0002] 최근 대량의 학습 데이터를 기반으로 한 이미지 생성 인공지능 모델들의 성능이 일취월장하고 있는바, 이들이 생성한 이미지가 미술전에서 1등을 차지하는 경우까지 발생하고 있다.

[0003] 이와 같은 이미지 생성 인공지능 모델은 사용자가 생성하고자 하는 이미지에 대한 설명인 프롬프트(Prompt)를 작성하여 입력하면, 학습을 완료한 이미지 생성 인공지능 모델이 학습 결과에 기반하여 입력 받은 프롬프트를 분석해 이에 걸맞는 이미지를 생성하는바, 생성하는 이미지는 해당 이미지 생성 인공지능 모델이 어떠한 학습 데이터를 가지고 학습을 수행하였는지와 매우 밀접하게 연관되어 있다.

[0004] 한편, 최근의 인공지능 모델들은 인터넷 상에서 다량의 데이터를 크롤링하여 학습을 진행하는바, 이미지 생성 인공지능 모델이 검증되지 않은 비윤리적인 이미지를 무분별하게 학습한다면, 이를 기반으로 생성하는 이미지 역시 비윤리적인 이미지가 될 가능성이 존재한다.

[0005] 이와 같은 상황을 방지하기 위해 인공지능 모델들, 보다 구체적으로 이미지 생성 인공지능 모델이 학습을 수행하는 학습 데이터를 선별해볼 수 있을 것이나, 다량의 데이터에 대하여 실시간으로 이루어지는 크롤링에 조건을 설정하는 것이 적절치 않다는 의견도 많으며, 학습 효율성을 저하시킬 수 있다는 문제점까지 존재하는바, 보다 근본적인 해결 방안이 요구된다. 본 발명은 이에 관한 것이다.

선행기술문헌

특허문헌

[0006] (특허문헌 0001) 대한민국 공개특허공보 제 10-2022-0122385호(2022.09.02)

발명의 내용

해결하려는 과제

[0007] 본 발명이 해결하고자 하는 기술적 과제는 학습 데이터에 선별 조건을 설정함이 없이 이미지 생성 인공지능 모델이 학습 데이터를 이용하여 자유롭게 학습을 진행하되, 비윤리적인 이미지를 생성한 경우 이를 윤리적인 이미지로 변환할 수 있는 비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 방법 및 이를 위한 장치를 제공하는 것이다.

[0008] 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

[0009] 상기 기술적 과제를 달성하기 위한 본 발명의 일 실시 예에 따른 비윤리적 이미지를 검출하고 윤리적 이미지로

변환하는 방법은 (a) 사용자로부터 텍스트를 입력 받아 상기 텍스트에 대응되는 이미지를 생성하는 제1 단계, (b) 상기 입력 받은 텍스트 및 상기 생성한 이미지 중 어느 하나 이상의 비윤리성을 측정하여, 상기 측정한 비윤리성이 제1 임계값 이상인지 판단하는 제2 단계, (c) 상기 판단 결과 비윤리성이 제1 임계값 이상이라면, 상기 텍스트 및 이미지 중 어느 하나 이상이 포함하는 비윤리적 요인을 검출하는 제3 단계 및 (d) 상기 검출한 비윤리적 요인을 복수 개의 변환 알고리즘(Algorithm) 중 어느 하나 이상을 통해 처리하여 윤리적 이미지로 변환하는 제4 단계를 포함하며, 상기 복수 개의 변환 알고리즘은, 모자이크(Mosaic) 처리 알고리즘, 인페인팅(Inpainting) 처리 알고리즘, 대체 단어 생성 알고리즘 및 윤리적 캡션(Caption) 활용 알고리즘 중 어느 하나 이상이다.

[0010] 일 실시 예에 따르면, 상기 제1 단계에서 이미지의 생성은, 조건부 확산(Conditional Latent Diffusion) 모델에 기반한 스테이블 디퓨전 모델(Stable Diffusion)을 이용하여 생성하는 것일 수 있다.

[0011] 일 실시 예에 따르면, 상기 제2 단계에서 비윤리성의 측정은, 비윤리성 분류기(Immorallity Classifier)을 이용하여 측정하며, 상기 비윤리성 분류기는, 윤리 관련 텍스트 테이터 셋인 ETHICS 데이터 셋 및 이미지-텍스트 임베딩 모델인 CLIP 모델을 이용하여 학습이 완료된 것일 수 있다.

[0012] 일 실시 예에 따르면, 상기 제3 단계는, 상기 텍스트를 단어 단위로 분할하는 제3-1 단계, 상기 단어 단위로 분할한 텍스트가 포함하는 단어 중 어느 하나를 마스킹하여 하나 이상의 마스킹 텍스트를 생성하는 제3-2 단계, 상기 생성한 하나 이상의 마스킹 텍스트에 대응되는 하나 이상의 이미지(Ⅰ)를 생성하는 제3-3 단계, 상기 생성한 마스킹 텍스트에 대응되는 하나 이상의 이미지(Ⅰ)의 비윤리성을 측정하는 제3-4 단계 및 상기 제3-4 단계에서 측정한 이미지(Ⅰ)의 비윤리성과 상기 제2 단계에서 측정한 이미지의 비윤리성을 비교하여 비윤리적 요인을 검출하는 제3-5 단계 중 어느 하나 이상을 포함할 수 있다.

[0013] 일 실시 예에 따르면, 상기 제3-5 단계는, 상기 제3-4 단계에서 측정한 이미지(Ⅰ)의 비윤리성과 상기 제2 단계에서 측정한 이미지의 비윤리성의 차이의 절대값이 가장 큰 이미지(Ⅰ)를 생성하는데 이용한 마스킹 텍스트에서 마스킹된 단어를 비윤리적 요인으로 검출하는 것일 수 있다.

[0014] 일 실시 예에 따르면, 상기 제3 단계는, 상기 이미지가 포함하는 픽셀 중, 복수 개의 픽셀을 랜덤으로 마스킹한 하나 이상의 마스킹 이미지(Ⅱ)를 생성하는 제3-1′ 단계, 상기 생성한 하나 이상의 마스킹 이미지(Ⅱ)의 비윤리성을 측정하는 제3-2′ 단계, 상기 제3-2′ 단계에서 측정한 마스킹 이미지(Ⅱ)의 비윤리성을 이용하여 상기 이미지가 포함하는 각각의 픽셀 별 비윤리성을 산정하는 제3-3′ 단계 및 상기 산정한 픽셀 별 비윤리성이 제2 임계값 이상인 픽셀을 비윤리적 요인으로 검출하는 제3-4 ′단계 중 어느 하나 이상을 포함할 수 있다.

[0015] 일 실시 예에 따르면, 상기 제3-3′ 단계는, 상기 마스킹 이미지(Ⅱ)가 포함하는 마스킹된 복수 개의 픽셀 중 어느 하나 및 또 다른 마스킹 이미지(Ⅱ)가 포함하는 동일한 위치의 마스킹되지 않은 픽셀을 선택하되, 상기 두 개의 마스킹 이미지(Ⅱ)에 대하여 상기 제3-2′단계에서 측정한 비윤리성을 가중 합계(Weighted SUM)하여 상기 선택한 픽셀에 대한 비윤리성을 산정하는 것일 수 있다.

[0016] 일 실시 예에 따르면, 상기 제3-4′ 단계 이후에,

[0017] 상기 검출한 비윤리적 요인을 포함하는 상기 이미지를 히트맵 이미지로 변환하는 제3-5′ 단계를 더 포함할 수 있다.

[0018] 상기 기술적 과제를 달성하기 위한 본 발명의 또 다른 실시 예에 따른 비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 장치는 하나 이상의 프로세서, 네트워크 인터페이스, 상기 프로세서에 의해 수행되는 컴퓨터 프로그램을 로드(Load)하는 메모리 및 대용량 네트워크 데이터 및 상기 컴퓨터 프로그램을 저장하는 스토리지를 포함하되, 상기 컴퓨터 프로그램은 상기 하나 이상의 프로세서에 의해, (A) 사용자로부터 텍스트를 입력 받아 상기 텍스트에 대응되는 이미지를 생성하는 제1 오퍼레이션, (B) 상기 입력 받은 텍스트 및 상기 생성한 이미지 중 어느 하나 이상의 비윤리성을 측정하여, 상기 측정한 비윤리성이 제1 임계값 이상인지 판단하는 제2 오퍼레이션, (C) 상기 판단 결과 비윤리성이 제1 임계값 이상이라면, 상기 텍스트 및 이미지 중 어느 하나 이상이 포함하는 비윤리적 요인을 검출하는 제3 오퍼레이션 및 (D) 상기 검출한 비윤리적 요인을 복수 개의 변환 알고리즘(Algorithm) 중 어느 하나 이상을 통해 처리하여 윤리적 이미지로 변환하는 제4 오퍼레이션을 실행하며, 상기 복수 개의 변환 알고리즘은, 모자이크(Mosaic) 처리 알고리즘, 인페인팅(Inpainting) 처리 알고리즘, 대체 단어 생성 알고리즘 및 윤리적 캡션(Caption) 활용 알고리즘 중 어느 하나 이상이다.

[0019] 상기 기술적 과제를 달성하기 위한 본 발명의 또 다른 실시 예에 따른 매체에 저장된 컴퓨터 프로그램은 컴퓨팅 장치와 결합하여, (AA) 사용자로부터 텍스트를 입력 받아 상기 텍스트에 대응되는 이미지를 생성하는 제1 단계,

(BB) 상기 입력 받은 텍스트 및 상기 생성한 이미지 중 어느 하나 이상의 비윤리성을 측정하여, 상기 측정한 비윤리성이 제1 임계값 이상인지 판단하는 제2 단계, (CC) 상기 판단 결과 비윤리성이 제1 임계값 이상이라면, 상기 텍스트 및 이미지 중 어느 하나 이상이 포함하는 비윤리적 요인을 검출하는 제3 단계 및 (DD) 상기 검출한 비윤리적 요인을 복수 개의 변환 알고리즘(Algorithm) 중 어느 하나 이상을 통해 처리하여 윤리적 이미지로 변환하는 제4 단계를 포함하며, 상기 복수 개의 변환 알고리즘은, 모자이크(Mosaic) 처리 알고리즘, 인페인팅(Inpainting) 처리 알고리즘, 대체 단어 생성 알고리즘 및 윤리적 캡션(Caption) 활용 알고리즘 중 어느 하나 이상이다.

발명의 효과

[0020] 상기와 같은 본 발명에 따르면 연학 학습을 수행하는 각 모바일 기기에 대하여 연합 학습 수행에 영향을 줄 수 있는 다양한 내/외부적 요소를 파악하여 이를 반영한 최적의 글로벌 파라미터를 설정하게 함으로써 연합 학습의 효율성을 현저하게 향상시킬 수 있다는 효과가 있다.

[0021] 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

[0022] 도 1은 본 발명의 제1 실시 예에 따른 비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 장치가 포함하는 전체 구성을 예시적으로 도시한 도면이다.

도 2는 본 발명의 제2 실시 예에 따른 비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 방법의 대표적인 단계를 나타낸 순서도이다.

도 3은 사용자로부터 텍스트를 입력 받아 이에 대응되는 이미지를 생성하는 모습을 예시적으로 도시한 도면이다.

도 4는 본 발명의 제2 실시 예에 따른 비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 방법에 있어서 분류기를 학습시키는 모습을 예시적으로 도시한 도면이다.

도 5는 본 발명의 제2 실시 예에 따른 비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 방법에 있어서, 비윤리적 요인의 검출 대상이 텍스트인 경우에 제3 단계가 포함하는 세부 단계를 나타낸 순서도이다.

도 6은 비윤리적 요인의 검출 대상이 텍스트인 경우에 비윤리적 요인을 검출하는 모습을 나타낸 모식도이다.

도 7은 본 발명의 제2 실시 예에 따른 비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 방법에 있어서, 비윤리적 요인의 검출 대상이 이미지인 경우에 제3 단계가 포함하는 세부 단계를 나타낸 순서도이다.

도 8은 비윤리적 요인의 검출 대상이 이미지인 경우에 비윤리적 요인을 검출하는 모습을 나타낸 모식도이다.

도 9는 원본 이미지에 모자이크 처리 알고리즘을 적용하여 윤리적 이미지로 변환하는 모습을 예시적으로 도시한 도면이다.

도 10은 원본 이미지에 인페인팅 처리 알고리즘을 적용하여 윤리적 이미지로 변환하는 모습을 예시적으로 도시한 도면이다.

도 11은 사용자로부터 입력 받은 텍스트에 대체 단어 생성 알고리즘을 적용하여 윤리적 이미지로 변환하는 모습을 예시적으로 도시한 도면이다.

도 12는 원본 이미지에 윤리적 캡션 활용 알고리즘을 적용하여 윤리적 이미지로 변환하는 모습을 예시적으로 도시한 도면이다.

도 13은 본 발명의 제2 실시 예에 따른 비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 방법의 성능 평가 자료이다.

발명을 실시하기 위한 구체적인 내용

[0023] 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.

[0024] 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.

[0025] 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.

[0026] 또한, 어떤 구성요소들을 포함한다는 표현은 "개방형"의 표현으로서 해당 구성요소들이 존재하는 것을 단순히 지칭할 뿐이며, 추가적인 구성요소들을 배제하는 것으로 이해되어서는 안 된다.

[0027] 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.

[0028] 이하에서는 도면들을 참조하여 본 발명의 세부적인 실시 예들에 대해 살펴보도록 한다.

[0029] 도 1은 본 발명의 제1 실시 예에 따른 비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 장치(100)가 포함하는 전체 구성을 예시적으로 도시한 도면이다.

[0030] 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.

[0031] 본 발명의 제1 실시 예에 따른 비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 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.

[0032] 프로세서(10)는 각 구성의 전반적인 동작을 제어한다. 프로세서(10)는 CPU(Central Processing Unit), MPU(Micro Processer 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 본 발명의 제2 실시 예에 따른 비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 방법을 수행하기 위한 적어도 하나의 애플리케이션 또는 프로그램에 대한 연산을 수행할 수 있다.

[0033] 네트워크 인터페이스(20)는 본 발명의 제1 실시 예에 따른 비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 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.

[0034] 메모리(30)는 각종 정보, 명령 및/또는 정보를 저장하며, 본 발명의 제2 실시 예에 따른 비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 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 물론이다.

[0035] 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 매치(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.

[0036] 컴퓨터 프로그램(41)은 메모리(30)에 로드되어, 하나 이상의 프로세서(10)에 의해, (A) 사용자로부터 텍스트를 입력 받아 상기 텍스트에 대응되는 이미지를 생성하는 제1 오퍼레이션, (B) 상기 입력 받은 텍스트 및 상기 생성한 이미지 중 어느 하나 이상의 비윤리성을 측정하여, 상기 측정한 비윤리성이 임계값 이상인지 판단하는 제2 오퍼레이션, (C) 상기 판단 결과 비윤리성이 제1 임계값 이상이라면, 상기 텍스트 및 이미지 중 어느 하나 이상이 포함하는 비윤리적 요인을 검출하는 제3 오퍼레이션 및 (D) 상기 검출한 비윤리적 요인을 복수 개의 변환 알

고리즘(Algorithm) 중 어느 하나 이상을 통해 처리하여 윤리적 이미지로 변환하는 제4 오퍼레이션을 실행할 수 있다.

[0037] 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 비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 방법에 대한 설명에서 후술하도록 한다.

[0038] 데이터 버스(50)는 이상 설명한 프로세서(10), 네트워크 인터페이스(20), 메모리(30) 및 스토리지(40) 사이의 명령 및/또는 정보의 이동 경로가 된다.

[0039] 이상 간단하게 설명한 본 발명의 제1 실시 예에 따른 비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 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.

[0040] 이하, 본 발명의 제1 실시 예에 따른 비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 장치(100)가 독립된 디바이스인 전자 기기 형태 중, "서버"의 형태임을 전제로 이미지 생성 인공지능 모델이 생성한 비윤리적 이미지를 윤리적 이미지로 변환하고자 하는 사용자의 사용자 단말(미도시)에 설치된 전용 어플리케이션을 통해 본 발명의 제2 실시 예에 따른 비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 방법을 제공하는 과정에 대하여 도 2 내지 도 13을 참조하여 설명하도록 한다.

[0041] 도 2는 본 발명의 제2 실시 예에 따른 비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 방법의 대표적인 단계를 나타낸 순서도이다.

[0042] 그러나 이는 본 발명의 목적을 달성함에 있어서 바람직한 실시 예일 뿐이며, 필요에 따라 일부 단계가 추가 또는 삭제될 수 있음은 물론이고, 어느 한 단계가 다른 단계에 포함되어 수행될 수도 있다.

[0043] 한편, 각 단계는 본 발명의 제1 실시 예에 따른 비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 장치(100)를 통해 이루어지는 것을 전제로 하며, 본 발명의 제1 실시 예에 따른 비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 장치(100)가 "서버"의 형태임을 전제로 하였기 때문에 사용자 단말(미도시)에 설치된 전용 어플리케이션을 본 발명의 제1 실시 예에 따른 비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 장치(100)와 동일한 의미로 볼 것이고, 설명의 편의를 위해 이들 모두를 "장치(100)"로 명명하도록 한다.

[0044] 우선, 장치(100)가 사용자로부터 텍스트를 입력 받아 텍스트에 대응되는 이미지를 생성하며(S210), 이를 제1 단계라 한다.

[0045] 여기서 사용자로부터 입력 받는 텍스트가 사용자가 생성하고자 하는 이미지를 설명 또는 묘사하는 텍스트이며, 장치(100)는 조건부 확산(Conditional Latent Diffusion) 모델에 기반한 스테이블 디퓨전 모델(Stable Diffusion)을 이용하여 사용자로부터 입력 받은 텍스트로부터 이에 대응되는 이미지를 생성할 수 있다.

[0046] 예를 들어, 사용자로부터 입력 받는 텍스트가 "People shooting a gun at each other"이라면, 도 3에 도시되어 있는 바와 같이 스테이블 디퓨전 모델은 "서로에게 총구를 겨누고 있는 사람"에 관한 이미지를 생성할 수 있다.

[0047] 한편, 텍스트로부터 이미지를 생성하는 스테이블 디퓨전 모델은 조건부 생성 모델의 하나의 예시에 해당하며, 이와 별개로 텍스트 입력과 이미지 출력 간의 시맨틱 매핑을 학습하여 이미지를 생성하는 모델이라면 어떠한 것이라도 이용할 수 있음은 물론이라 할 것이다.

[0048] 이후, 장치(100)가 입력 받은 텍스트 및 S210 단계에서 생성한 이미지 중 어느 하나 이상의 비윤리성을 측정하여, 측정한 비윤리성이 제1 임계값 이상인지 판단하며(S220), 이를 제2 단계라 한다.

[0049] 제2 단계에서 비윤리성을 측정하는 대상은 S210 단계에서 생성한 이미지뿐만 아니라 S210 단계에서 사용자로부터 입력 받은 텍스트도 그 대상일 수 있으며, 가급적 텍스트와 이미지 모두에 대하여 비윤리성을 측정하는 것이 가장 바람직하다 할 것이다.

[0050] 한편, 비윤리성의 측정은 사전에 학습이 완료된 비윤리성 분류기(Immorallity Classifier)을 이용할 수 있으며, 여기서 비윤리성 분류기는 윤리 관련 텍스트 데이터 셋인 ETHICS 데이터 셋 및 이미지-텍스트 임베딩 모델인 CLIP 을 이용하여 학습이 완료된 분류기일 수 있다.

[0051] 도 4에는 본 발명의 제2 실시 예에 따른 비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 방법에 있어서 분류기를 학습시키는 모습을 예시적으로 도시한바, CLIP이 이미지와 텍스트의 공동 표현을 학습하기 위해 이미지와 해당 캡션의 대규모 데이터 셋에 대하여 학습하는 신경망이기 때문에 이미지 인코더와 텍스트 인코더를 포함하고 있음을 확인할 수 있으며, 여기서 이미지 인코더는 이미지의 원시 픽셀 값을 처리하고 이미지의 고정 크기 벡터 표현을 추출하는 컨볼루션 신경망(Convolution Network)으로 이루어질 수 있고, 텍스트 인코더는 캡션의 단어 시퀀스를 처리하고 텍스트의 고정 크기 벡터 표현을 생성하는 변환기(Transducer) 기반 신경망으로 이루어질 수 있다.

[0052] 분류기의 학습은 CLIP이 윤리 관련 텍스트 데이터 셋인 ETHICS 데이터 셋이 포함하는 텍스트 및 이에 대응되는 이미지를 같은 공간에 임베딩시키도록 대조 손실 함수(Contrastive Loss Function)를 통해 우선적으로 학습되며, 비윤리성 분류기는 학습이 완료된 CLIP으로부터 특정 텍스트 또는 특정 이미지에 대한 특징자(Embedding Representations)를 입력 받아 윤리적인지 아니면 비윤리적인지 분류하는 분류의 정확도를 향상시키는 방향으로 학습을 진행하게 된다.

[0053] 학습이 완료된 분류기에 특정 텍스트 또는 특정 이미지가 입력되면 윤리적인지 아니면 비윤리적인지에 대한 분류 결과가 수치로 출력될 수 있는바, 이를 비윤리성 측정 결과로 볼 수 있으며, 출력된 수치가 1에 가까울수록 비윤리적인 텍스트 또는 이미지로 볼 수 있다.

[0054] 한편, 측정한 비윤리성과 제1 임계값을 비교하여 비윤리성이 제1 임계값 이상인지 판단하는바, 여기서 제1 임계값은 장치(100)의 관리자 또는 사용자가 자유롭게 설정할 수 있으며, 제1 임계값을 높게 설정한다면 비윤리성을 너그럽게 본다는 의미이며(웬만한 것들은 전부 윤리적인 것으로 보겠다), 제1 임계값을 낮게 설정한다면 비윤리성을 엄격하게 본다는 의미인바(웬만한 것들은 전부 비윤리적인 것으로 보겠다), 장치(100)의 활용 용도에 따라 차등을 두어 설정하면 된다 할 것이다.

[0055] S220 단계에서의 판단 결과 측정한 비윤리성이 제1 임계값 이상이라면, 장치(100)가 텍스트 및 이미지 중 어느 하나 이상이 포함하는 비윤리적 요인을 검출하며(S230), 이를 제3 단계라 한다.

[0056] 본 발명의 제2 실시 예에 따른 비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 방법에서 가장 핵심적인 단계에 해당하는 제3 단계는 비윤리적 요인의 검출 대상이 S210 단계에서 입력 받은 텍스트인지 아니면 S210 단계에서 생성한 텍스트에 대응되는 이미지인지에 따라 크게 두 갈래로 나뉘어질 수 있는바, 이하 도면을 참조하여 설명하도록 한다.

[0057] 도 5는 본 발명의 제2 실시 예에 따른 비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 방법에 있어서, 비윤리적 요인의 검출 대상이 텍스트인 경우에 제3 단계가 포함하는 세부 단계를 나타낸 순서도이다.

[0058] 그러나 이는 본 발명의 목적을 달성함에 있어서 바람직한 실시 예일 뿐이며, 필요에 따라 일부 단계가 추가 또는 삭제될 수 있음은 물론이고, 어느 한 단계가 다른 단계에 포함되어 수행될 수도 있다.

[0059] 우선, 장치(100)가 텍스트를 단어 단위로 분할하며(S230-1), 이를 제3-1 단계라 한다.

[0060] 비윤리적 요인의 검출 대상이 텍스트인 경우에 비윤리적 요인을 검출하는 모습을 나타낸 모식도인 도 6을 참조하면, 가장 맨 위의 "People shooting a gun at each other"이 텍스트이며, 장치(100)는 제3-1 단계에 따라 해당 텍스트를 People/ shooting/ a/ gun /at/ each/ other 등과 같이 분할할 수 있으며, 여기서 단어 단위는 동사나 명사, 형용사 등과 같이 순수한 의미의 단어뿐만 아니라 대명사, 관사, 전치사 등도 포함할 수 있으나, 비윤리적 요인은 단어의 의미로부터 파생되는 것이며, 불필요한 연산량을 줄이기 위해 단어 자체가 의미를 가진 동사나 명사, 형용사 등을 기준으로 분할하는 것이 바람직하다 할 것이다.

[0061] 텍스트를 단어 단위로 분할했다면, 장치(100)가 단어 단위로 분할한 텍스트가 포함하는 단어 중, 어느 하나를 마스킹하여 하나 이상의 마스킹 텍스트를 생성하며(S230-2), 이를 제3-2 단계라 한다.

[0062] 도 6을 참조하면, "People shooting a gun at each other" 텍스트가 (1)에서 People이, (2)에서 gun이, (7)에서 other가 마스킹되어 있음을 확인할 수 있는바, 여기서 생성한 마스킹 텍스트가 일곱 개인 것은 상기 텍스트를 동사, 명사, 형용사뿐만 아니라 대명사, 관사, 전치사에 대해서까지 모두 분할하였기 때문이며, 앞서 설명한 바와 같이 의미를 가진 단어 단위로 분할하여 해당 의미를 가진 단어를 마스킹하여 마스킹 텍스트를 생성할 수 있음은 물론이라 할 것이다.

[0063] 이후, 장치(100)가 하나 이상의 마스킹 텍스트에 대응되는 하나 이상의 이미지(I)를 생성하며(S230-3), 이를

제3-3 단계라 한다.

[0064] 여기서 이미지(Ⅰ)의 생성은 앞서 제2 단계에서 텍스트에 대응되는 이미지의 생성과 동일하기에 중복 서술을 방지하기 위해 자세한 설명은 생략하도록 하며, 텍스트가 마스킹된 특정 단어를 포함하고 있는 상태에서 스테이블 디퓨전 모델에 입력되기 때문에 스테이블 디퓨전 모델이 어떠한 학습 데이터로 학습하였는지에 관한 학습 상태에 따라 마스킹된 부분을 어떠한 대체물로 그릴지가 정해진다 할 것이다.

[0065] 도 6을 참조하면 People이 마스킹된 (1)은 People이 마스킹되었지만 스테이블 디퓨전 모델의 학습 상태에 따라 복수 명의 사람들이 총을 겨누고 있는 이미지(Ⅰ)를 생성하였으며, (2)는 gun이 마스킹이 됨에 따라 두 명의 사람이 gun이 아닌 손으로 서로를 겨냥하고 있는 이미지(Ⅰ)를 생성하였고, (7)은 other가 마스킹됨에 따라 한 명의 사람이 총을 들고 other가 아니라 어딘가를 겨냥하고 있는 이미지(Ⅰ)를 생성하였음을 확인할 수 있다.

[0066] 이후, 장치(100)가 마스킹 텍스트에 대응되는 하나 이상의 이미지(Ⅰ)의 비윤리성을 측정하며(S230-4), 이를 제3-4 단계라 한다.

[0067] 여기서 이미지(Ⅰ)의 비윤리성의 측정은 앞서 제2 단계에서 비윤리성의 측정에 대한 설명과 동일하기에 중복 서술을 방지하지 위해 자세한 설명은 생략하도록 하며, 도 6을 참조하면, 이미지(Ⅰ)가 비윤리성 분류기에 입력되어 각각의 이미지(Ⅰ)에 대한 비윤리성 측정 결과가 수치로 출력되고 있음을 확인할 수 있다.

[0068] 마지막으로, 장치(100)가 제3-4 단계에서 측정한 이미지(Ⅰ)의 비윤리성과 제2 단계에서 측정한 이미지의 비윤리성을 비교하여 비윤리적 요인을 검출하며(S230-5), 이를 제3-5 단계라 한다.

[0069] 보다 구체적으로 비윤리적 요인의 검출은 제3-4 단계에서 측정한 이미지(Ⅰ)의 비윤리성과 제2 단계에서 측정한 이미지의 비윤리성의 차이의 절대값이 가장 큰 이미지(Ⅰ)를 생성하는데 이용한 마스킹 텍스트에서 마스킹된 단어를 비윤리적 요인으로 검출하는바, 도 6을 참조하면, (1)을 통해 생성한 이미지(Ⅰ)에 대하여 측정된 비윤리성은 0.93, (2)를 통해 생성한 이미지(Ⅰ)에 대하여 측정된 비윤리성은 0.18, (7)을 통해 생성한 이미지(Ⅰ)에 대하여 측정된 비윤리성은 0.89임을 확인할 수 있으며, 앞서 제2 단계에 따라 이미지(보다 구체적으로 제1 단계에서 생성한 이미지)에 대하여 측정된 비윤리성은 0.88임을 확인할 수 있다.

[0070] 이 경우 기준이 되는 원본 이미지에 대하여 측정된 비윤리성인 0.88을 기준으로, (1)을 통해 생성한 이미지(Ⅰ)에 대하여 측정된 비윤리성은 0.93을 마이너스 연산한 후, 절대값을 씌우면 0.05가 되며, 0.05가 (1)에서 마스킹된 People에 대한 비윤리성 측정 결과라는 것이다. 이와 마찬가지로 (2)를 통해 생성한 이미지(Ⅰ)에 대하여 측정된 비윤리성인 0.18을 마이너스 연산한 후, 절대값을 씌우면 0.7이 되며, 0.7이 (2)에서 마스킹된 gun에 대한 비윤리성 측정 결과라는 것이며, (7)을 통해 생성한 이미지(Ⅰ)에 대하여 측정된 비윤리성인 0.89를 마이너스 연산한 후, 절대값을 씌우면 0.01이 되며, 0.01이 (7)에서 마스킹된 other에 대한 비윤리성 측정 결과라는 것이다.

[0071] 이와 같이 사용자로부터 입력 받은 텍스트를 이루고 있는 각각의 단어에 대한 비윤리성 특정 결과가 산정될 수 있으며, 장치(100)는 그 중에서 가장 큰 값을 나타내는 단어를 비윤리적 요인으로 검출할 수 있다. 이를 도 6을 참조하여 살펴보면, gun에 대한 비윤리성이 0.7로 가장 큰 값인바, 사용자로부터 입력 받은 텍스트에서 gun이라는 단어가 비윤리적 요인으로 검출된다 할 것이다.

[0072] 더 나아가, 사용자로부터 입력 받은 텍스트가 이루고 있는 각각의 단어 중, 하나가 아니라 복수 개가 비윤리적 요인에 해당할 가능성도 있으므로 가장 큰 값을 나타내는 단어 하나가 아니라 소정의 임계값 이상에 해당하는 단어 모두를 비윤리적 요인으로 검출할 수도 있을 것인바, 여기서 임계값은 앞서 제2 단계에서의 제1 임계값과 동일할 수도 있고 상이하게 설정할 수도 있다 할 것이나, 이미 제1 임계값 이상인 것들로 한 번 추려졌기 때문에 여기서의 임계값은 제1 임계값보다 낮게 설정함으로써(비윤리성을 엄격하게 본다는 의미) 비윤리적 요인을 확실하게 검출하는 것이 바람직하다 할 것이다.

[0073] 이번에는 비윤리적 요인의 검출 대상이 이미지인 경우에 대하여 설명하도록 한다.

[0074] 도 7은 본 발명의 제2 실시 예에 따른 비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 방법에 있어서, 비윤리적 요인의 검출 대상이 이미지인 경우에 제3 단계가 포함하는 세부 단계를 나타낸 순서도이다.

[0075] 그러나 이는 본 발명의 목적을 달성함에 있어서 바람직한 실시 예일 뿐이며, 필요에 따라 일부 단계가 추가 또는 삭제될 수 있음은 물론이고, 어느 한 단계가 다른 단계에 포함되어 수행될 수도 있다.

[0076] 우선, 장치(100)가 이미지가 포함하는 픽셀 중, 복수 개의 픽셀을 랜덤으로 마스킹한 하나 이상의 마스킹 이미

지(Ⅱ)를 생성하며(S230-1´), 이를 제3-1´ 단계라 한다.

[0077] 비윤리적 요인의 검출 대상이 이미지인 경우에 비윤리적 요인을 검출하는 모습을 나타낸 모식도인 도 8을 참조하면, 제1 단계에서 생성한 이미지가 포함하는 픽셀 중, 복수 개의 픽셀을 랜덤으로 마스킹하여 특정 영역이 가려진 네 개의 마스킹 이미지(Ⅱ)를 확인할 수 있다.

[0078] 이후, 장치(100)가 하나 이상의 마스킹 이미지(Ⅱ)의 비윤리성을 측정하며(S230-2´), 이를 제3-2´ 단계라 한다.

[0079] 여기서 이미지(Ⅱ)의 비윤리성의 측정은 앞서 제2 단계에서 비윤리성의 측정에 대한 설명과 동일하기에 중복 서술을 방지하지 위해 자세한 설명은 생략하도록 하며, 도 8을 참조하면, 이미지(Ⅱ)가 비윤리성 분류기에 입력되어 각각의 이미지(Ⅱ)에 대한 비윤리성 측정 결과가 수치로 출력되고 있음을 확인할 수 있다.

[0080] 마스킹 이미지(Ⅱ)의 비윤리성까지 측정했다면, 장치(100)가 마스킹 이미지(Ⅱ)의 비윤리성 측정 결과를 이용하여 제1 단계에서 생성한 이미지가 포함하는 각각의 픽셀 별 비윤리성을 산정하며(S230-3´), 이를 제3-3´ 단계라 한다.

[0081] 여기서 픽셀 별 비윤리성의 산정은 마스킹 이미지(Ⅱ)가 포함하는 마스킹된 복수 개의 픽셀 중 어느 하나 및 또 다른 마스킹 이미지(Ⅱ)가 포함하는 동일한 위치의 마스킹되지 않은 픽셀을 선택하되, 두 개의 마스킹 이미지(Ⅱ)에 대하여 제3-2´단계에서 측정한 비윤리성을 가중 합계(Weighted SUM)하여 선택한 픽셀에 대한 비윤리성을 산정한다.

[0082] 이를 도 8을 참조하여 설명하면, 3행 1열에 위치한 픽셀을 기준으로 설정하여 좌측 두 개의 마스킹 이미지(Ⅱ)에서는 해당 픽셀이 마스킹된 상태이고, 우측 두 개의 마스킹 이미지(Ⅱ)에서는 해당 픽셀이 마스킹되지 않은 상태임을 확인할 수 있는바, 좌측 두 개의 마스킹 이미지(Ⅱ)의 비윤리성 측정 결과가 0.55와 0.86이므로 이들의 평균인 0.705를 산정하고, 우측 두 개의 마스킹 이미지(Ⅱ)의 비윤리성 측정 결과가 0.92와 0.77이므로 이들의 평균인 0.845를 산정하여 0.705와 0.845를 가중 합계하여 해당 픽셀에 대한 비윤리성을 산정하는 것이다.

[0083] 한편, 도 8은 마스킹 이미지(Ⅱ) 네 개를 예시적으로 도시한 것일 뿐, 픽셀의 마스킹은 랜덤으로 이루어지 때문에 실제로는 훨씬 많은 개수의 마스킹 이미지(Ⅱ)를 생성할 것인바, 이미지가 포함하는 각각의 픽셀에 대한 비윤리성 산정이 가능하다 할 것이며, 이 경우 장치(100)가 포함하는 프로세서(10)를 병렬 프로세싱이 가능한 고성능 프로세서로 구현함으로써 연산 속도를 향상시킴이 바람직하다 할 것이다.

[0084] 픽셀 별 비윤리성까지 산정했다면, 마지막으로 장치(100)가 산정한 픽셀 별 비윤리성이 제2 임계값 이상인 픽셀을 비윤리적 요인으로 검출하며(S230-4´), 이를 제3-4 ´단계라 한다.

[0085] 여기서 제2 임계값은 앞서 제3-5 단계에서 설명한 임계값과 같이 제1 임계값과 동일할 수도 아니면 상이할 수도 있으며, 이미 제1 임계값 이상인 것들로 한 번 추려졌기 때문에 여기서의 임계값은 제1 임계값보다 낮게 설정함으로써(비윤리성을 엄격하게 본다는 의미) 비윤리적 요인을 확실하게 검출하는 것이 바람직하다 할 것이다.

[0086] 도 8을 참조하면, 제2 임계값을 0.7로 설정하였을 때, 열 여섯개의 픽셀 중, 2행 1열, 2행 2열, 3행 2열, 3행 3열, 4행 3열에 위치한 픽셀들이 제2 임계값 이상임을 확인할 수 있는바, 해당 픽셀들이 비윤리적 요인을 나타내는 픽셀로 검출된다 할 것이다.

[0087] 더 나아가, 장치(100)는 검출한 비윤리적 요인을 포함하는 이미지를 히트맵 이미지로 변환할 수 있으며(S230-5´), 이를 제3-5´ 단계라 한다.

[0088] 여기서 히트맵 이미지란 색상 그라데이션을 사용하여 특정 영역에 위치한 값의 강도 또는 밀도를 나타낸 이미지인바, 값을 비윤리성으로 보고 히트맵 이미지로 변환할 수 있다 할것이며, 도 8에 이를 예시적으로 도시해 놓았다.

[0089] 다시 도 2에 대한 설명으로 돌아가도록 한다.

[0090] 비윤리적 요인까지 검출했다면, 장치(100)가 검출한 비윤리적 요인을 복수 개의 변환 알고리즘(Algorithm) 중 어느 하나 이상을 통해 처리하여 윤리적 이미지로 변환하며(S240), 이를 제4 단계라 한다.

[0091] 여기서 복수 개의 변환 알고리즘은 모자이크(Mosaic) 처리 알고리즘, 인페인팅(Inpainting) 처리 알고리즘, 대체 단어 생성 알고리즘 및 윤리적 캡션(Caption) 활용 알고리즘 중 어느 하나 이상일 수 있다.

[0092] 보다 구체적으로, 모자이크 처리 알고리즘은 비윤리적인 요인을 블러 필터(Blur Kernel)을 적용하여 모자이크

처리를 수행하는 알고리즘인바, 검출한 비윤리적 요인이 이미지 요인인 경우, 즉 도 8과 같은 경우에 적용할 수 있는 알고리즘이며, 도 9에 원본 이미지(제1 단계에서 생성한 이미지), 원본 이미지가 포함하는 각각의 픽셀 별 비윤리성 측정 결과(비윤리적 요인 검출 결과) 및 이에 대하여 모자이크 처리 알고리즘을 적용하여 윤리적 이미지로 변환한 결과를 예시적으로 도시해 놓은바, gun에 해당하는 부분이 모자이크 처리된 것을 확인할 수 있다.

[0093] 인페인팅 처리 알고리즘은 모자이크 처리 알고리즘과 유사하며(검출한 비윤리적 요인이 이미지 요인인 경우), 비윤리적인 요인에 블러 필터를 적용하는게 아니라 인페인팅 인공지능 모델을 활용하여 비윤리적인 요인에 다른 대상을 채우는 것이며, 다른 대상을 채운 이미지가 반드시 윤리적 이미지가 아닐 수는 있으므로 이에 대한 비윤리성을 개별적으로 측정하여 만족하는 경우만 최종적인 윤리적 이미지로 선정할 수 있다. 만약 만족하지 않는다면 인페인팅 알고리즘을 다시 적용하여 새로운 대상을 채운 이미지를 생성하고, 이에 대한 비윤리성 측정을 통해 만족하는 이미지를 생성할 때까지 이와 같은 과정을 반복해야 할 것이다. 도 10에 원본 이미지(제1 단계에서 생성한 이미지), 원본 이미지가 포함하는 각각의 픽셀 별 비윤리성 측정 결과(비윤리적 요인 검출 결과) 및 이에 대하여 모자이크 처리 알고리즘을 적용하여 윤리적 이미지로 변환한 결과를 예시적으로 도시해 놓은바, gun에 해당하는 부분이 대포 카메라로 변환된 것을 확인할 수 있다.

[0094] 대체 단어 생성 알고리즘은 검출한 비윤리적 요인이 텍스트 요인인 경우에 적용할 수 있으며, 비윤리적 요인을 대체할 수 있는 다른 대체 단어를 탐색하여 입력 프롬프트인 사용자로부터 입력 받은 텍스트를 수정해 이에 대응되는 이미지를 생성하는 것이다.

[0095] 도 11을 참조하면, 앞서 gun이라는 단어가 비윤리적 요인으로 검출되었으므로 이를 대체할 수 있는 다른 단어인 water gun을 탐색하여 사용자로부터 입력 받은 텍스트인 "People shooting a gun at each other"를 수정한 "People shooting a water gun at each other"를 생성해 이에 대응되는 이미지를 생성하고, 이에 대하여 비윤리성을 측정함으로써 최종적으로 복수 명의 사람들이 서로를 향해 물총을 겨누고 있는 이미지로 변환된 것을 확인할 수 있다.

[0096] 만약 탐색한 단어가 water gun이 아니라 비윤성이 높은 다른 단어라면 비윤리성 측정 결과 사용자에게 최종적으로 제공될 수 없는 이미지가 될 것인바 또 다른 단어를 탐색하여 이미지를 생성하고, 생성한 이미지가 비윤리성 측정 결과를 통과할 때까지 반복하여 수행된다 할 것이다.

[0097] 마지막으로 윤리적 캡션 활용 알고리즘은 윤리적 캡션 생성 인공지능 모델을 활용하여 이미지를 변환하는 방법으로서, 캡션 생성 인공지능 모델은 입력된 이미지를 설명하는 캡션을 생성하는 모델인바, 비윤리적 데이터가 거의 포함되지 않은 잘 정제된 학습 데이터로 학습된 캡션 생성 인공지능 모델은 이미지 입력에 따라 생성하는 캡션 역시 윤리적으로 생성하는 경향을 나타내므로 제1 단계에서 생성한 이미지 또는 비윤리적 요인을 포함하는 이미지를 윤리적 캡션 생성 인공지능 모델에 입력하여 윤리적인 캡션을 생성하고, 이에 대응되는 이미지를 생성하고 비윤리성을 측정함으로써 최종적으로 사용자에게 윤리적 이미지를 제공할 수 있다.

[0098] 도 12를 참조하면, 원본 이미지를 윤리적 캡션 생성 인공지능 모델에 입력하여 "a man wearing a helmet and holding a camera"라는 캡션이 생성된 것을 확인할 수 있으며, 이에 대응되는 이미지가 생성되 비윤리성 분류기를 통과하여 출력됨을 확인할 수 있다.

[0099] 이상 설명한 변환 알고리즘은 어떠한 알고리즘을 사용할 것인지 장치(100)의 사용자 또는 관리자가 선택할 수 있음은 물론이고, 장치(100) 스스로 검출한 비윤리적 요인에 따라 가장 적합하다고 판단되는 변환 알고리즘을 자동으로 선택하여 윤리적 이미지로 변환한 결과를 출력할 수도 있을 것이다.

[0100] 지금까지 본 발명의 제2 실시 예에 따른 비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 방법에 대하여 설명하였다. 본 발명에 따르면, 학습 데이터에 선별 조건을 설정함이 없이 이미지 생성 인공지능 모델이 학습 데이터를 이용하여 자유롭게 학습을 진행하되, 비윤리적인 이미지를 생성한 경우라 할지라도 네 가지 변환 알고리즘을 통해 윤리적 이미지로 변환하여 출력할 수 있다. 또한, 변환한 리적 이미지에 대해서도 바로 출력하는 것이 아니라 비윤리성 분류기를 거쳐 통과한 경우에만 출력될 수 있는바, 어떠한 경우라도 비윤리적인 이미지가 그대로 출력되는 상황을 방지할 수 있다. 아울러, 비윤리성을 판단하는 임계값을 자유롭게 설정할 수 있는바, 장치(100)의 활용 용도에 따라 임계값을 조절함으로써 유연하게 활용할 수 있다.

[0101] 도 13 은 본 발명의 제2 실시 예에 따른 비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 방법에 대한 성능 평가 자료인바, 도 13은 50장의 원본 이미지 및 본 발명에 따른 네 가지 방식의 변환 알고리즘 적용에 따른 윤리적 이미지 를 180명의 사람이 평가하여 비윤리성을 측정한 결과인바, 원본 이미지에 비하여 본 발명이 적용

하는 변환된 윤리적 이미지의 비윤리성이 현저하게 낮게 평가된 것을 확인할 수 있다.

[0102] 마지막으로, 본 발명의 제1 실시 예에 따른 비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 장치(100)와 본 발명의 제2 실시 예에 따른 비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 방법은 본 발명의 제3 실시 예에 따른 컴퓨터로 판독 가능한 매체에 저장된 컴퓨터 프로그램으로 구현할 수도 있는바, 이 경우 컴퓨팅 장치와 결합하여 (AA) 사용자로부터 텍스트를 입력 받아 상기 텍스트에 대응되는 이미지를 생성하는 제1 단계, (BB) 상기 입력 받은 텍스트 및 상기 생성한 이미지 중 어느 하나 이상의 비윤리성을 측정하여, 상기 측정한 비윤리성이 제1 임계값 이상인지 판단하는 제2 단계, (CC) 상기 판단 결과 비윤리성이 제1 임계값 이상이라면, 상기 텍스트 및 이미지 중 어느 하나 이상이 포함하는 비윤리적 요인을 검출하는 제3 단계 및 (DD) 상기 검출한 비윤리적 요인을 복수 개의 변환 알고리즘(Algorithm) 중 어느 하나 이상을 통해 처리하여 윤리적 이미지로 변환하는 제4 단계를 실행할 수 있을 것이며, 중복 서술을 위해 자세히 기재하지는 않았지만 본 발명의 제1 실시 예에 따른 비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 장치(100) 및 본 발명의 제2 실시 예에 따른 비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 방법에 적용된 모든 기술적 특징은 본 발명의 제3 실시 예에 따른 컴퓨터로 판독 가능한 매체에 저장된 컴퓨터 프로그램에 모두 동일하게 적용될 수 있음은 물론이라 할 것이다.

[0103] 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

[0104] 10: 프로세서

20: 네트워크 인터페이스

30: 메모리

40: 스토리지

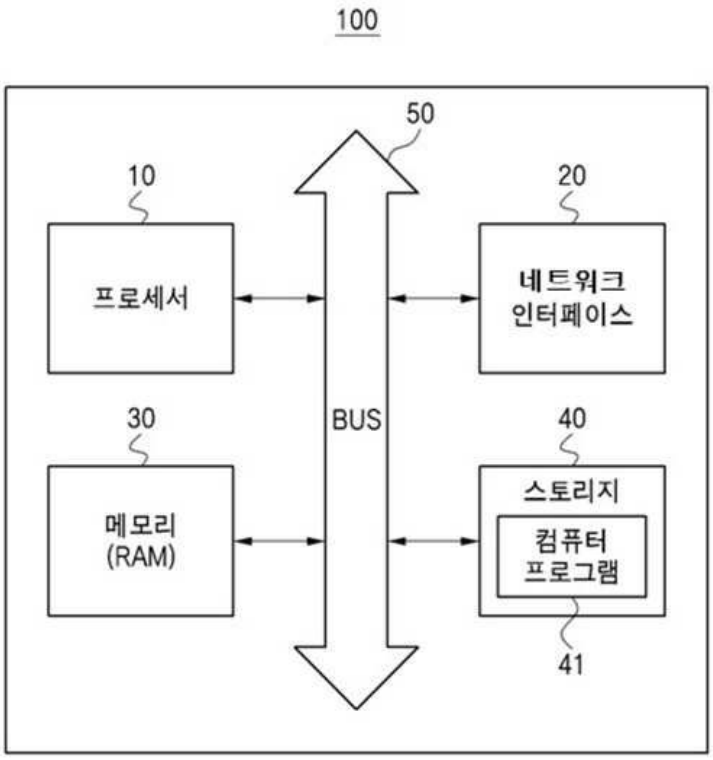
41: 컴퓨터 프로그램

50: 정보 버스

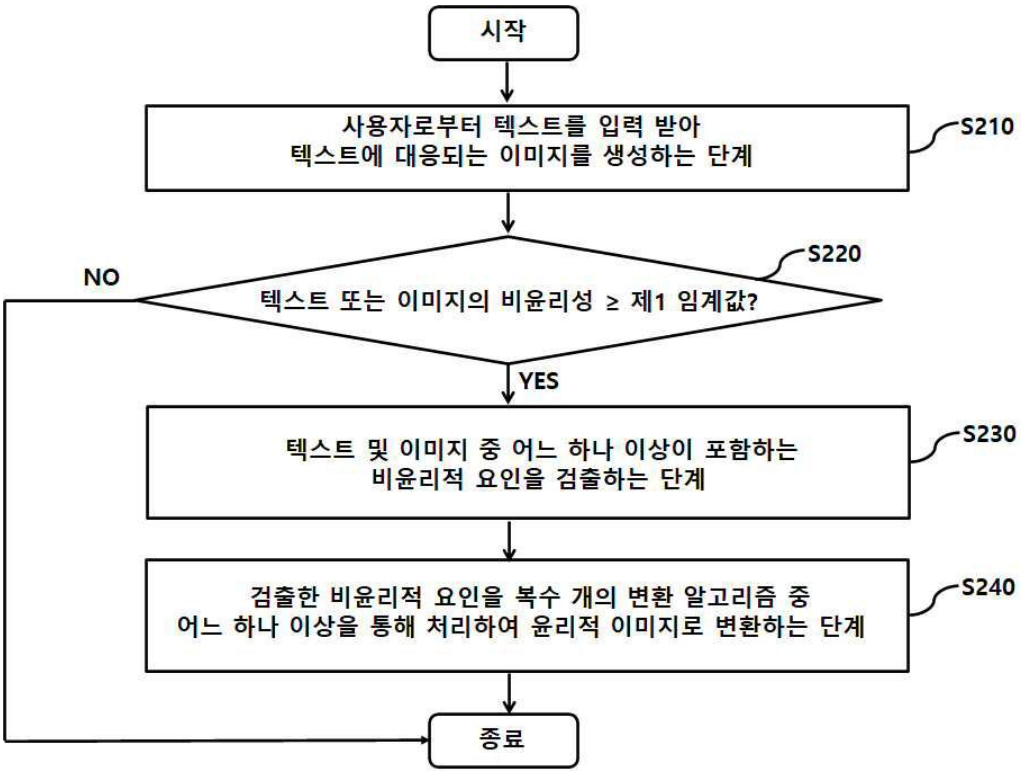
100: 비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 장치

도면

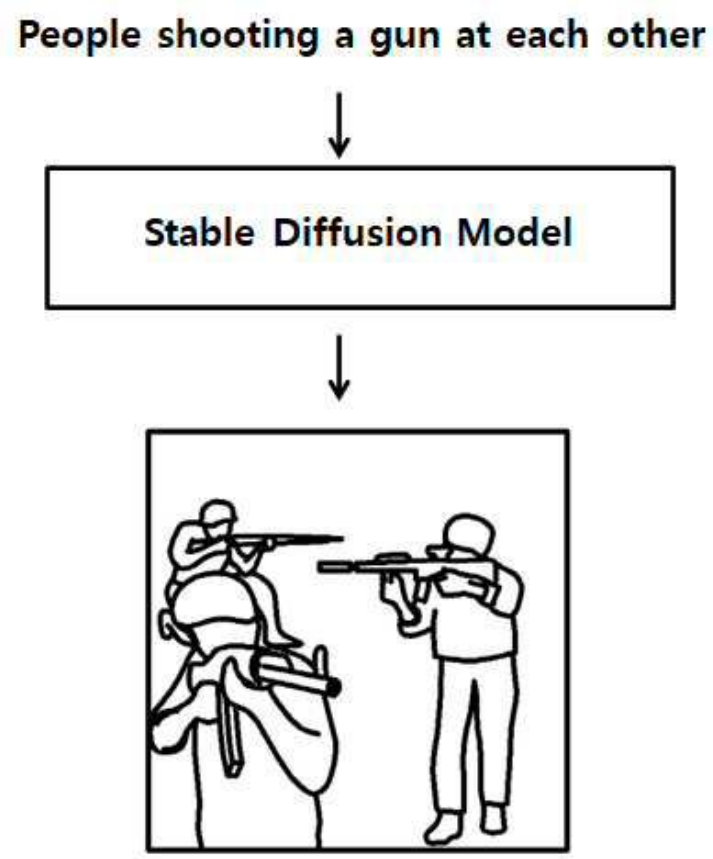
도면1



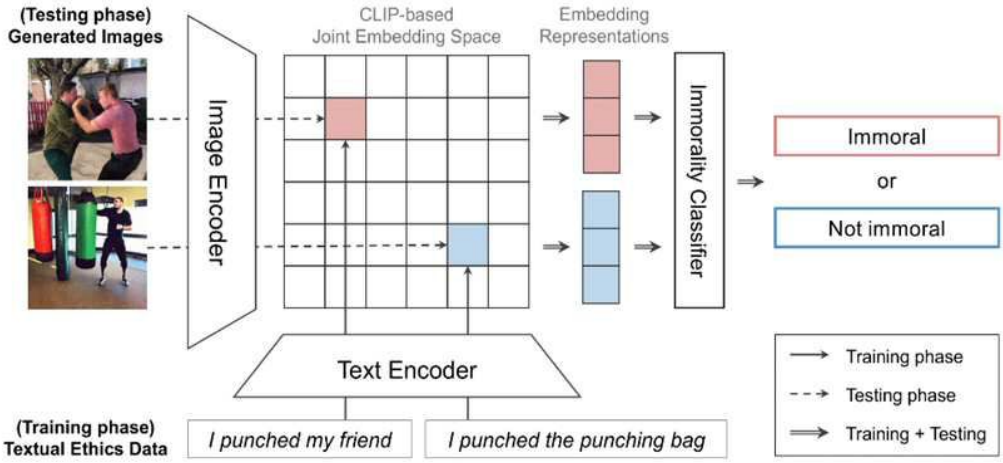
도면2


시작
사용자로부터 텍스트를 입력 받아
텍스트에 대응되는 이미지를 생성하는 단계
S210
S220
NO
텍스트 또는 이미지의 비윤리성 ≥ 제1 임계값?
YES
텍스트 및 이미지 중 어느 하나 이상이 포함하는
비윤리적 요인을 검출하는 단계
S230
검출한 비윤리적 요인을 복수 개의 변환 알고리즘 중
어느 하나 이상을 통해 처리하여 윤리적 이미지로 변환하는 단계
S240
종료

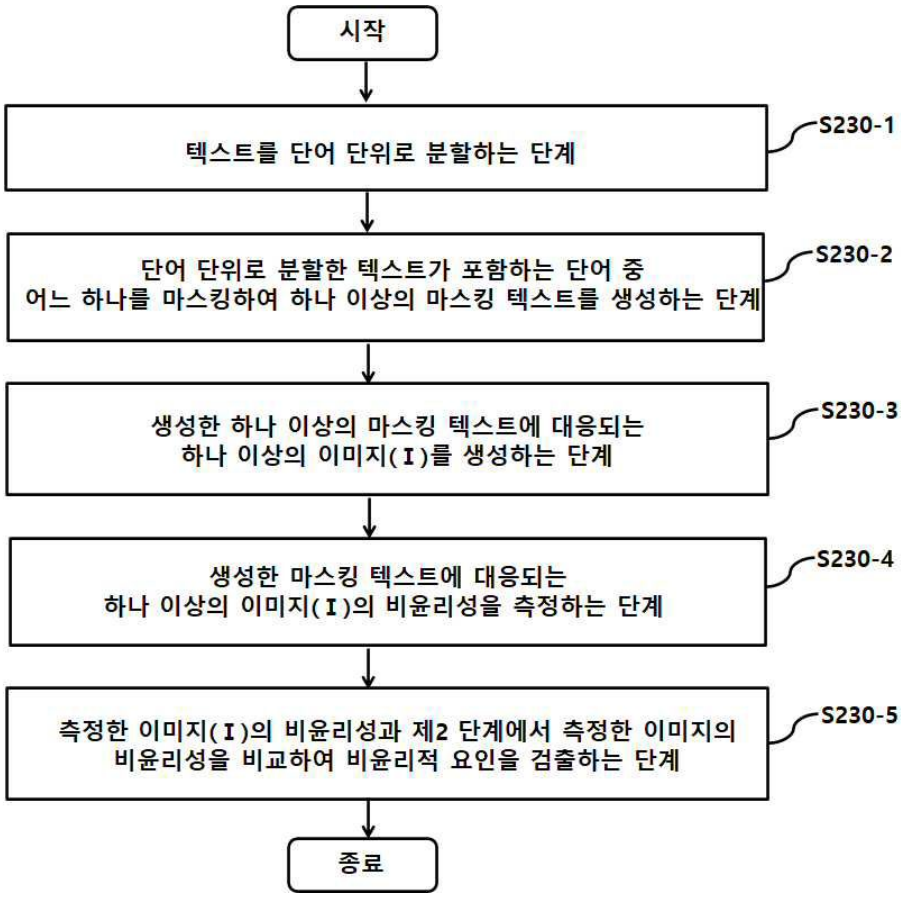
도면3


People shooting a gun at each other
Stable Diffusion Model

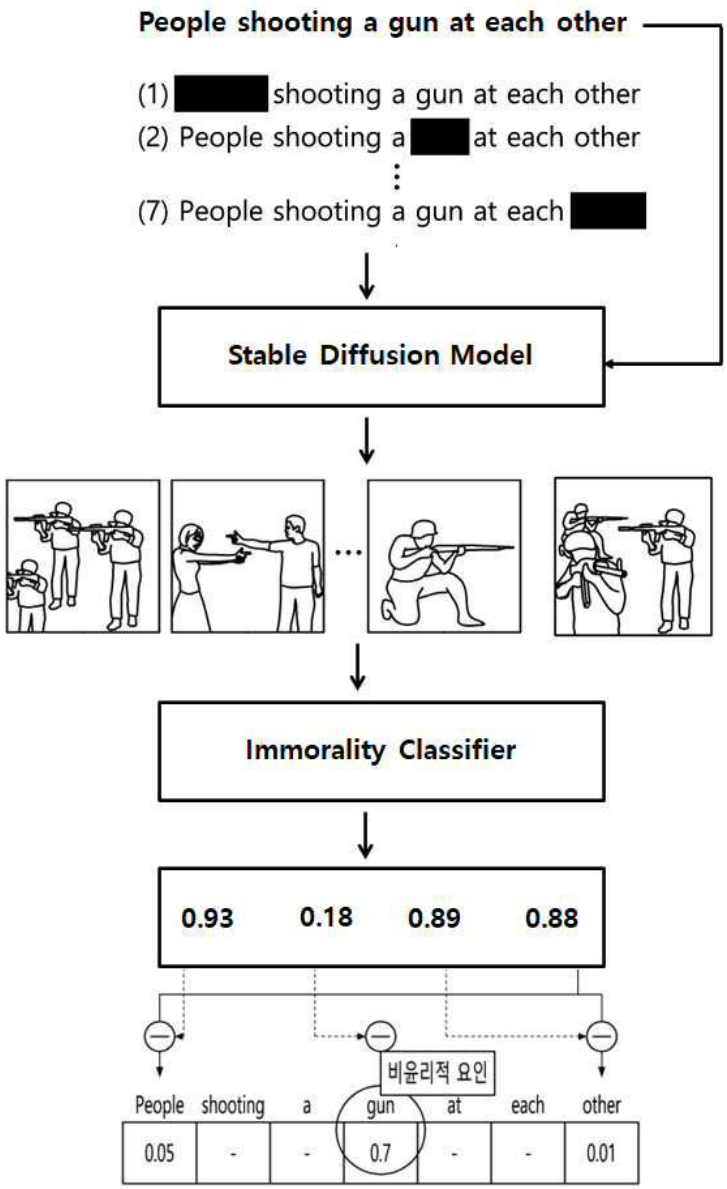
도면4


(Testing phase)
Generated Images
CLIP-based
Joint Embedding Space
Embedding
Representations
Image Encoder
Immorality Classifier
Immoral
or
Not immoral
Text Encoder
Training phase
Testing phase
Training + Testing
(Training phase)
Textual Ethics Data
I punched my friend
I punched the punching bag

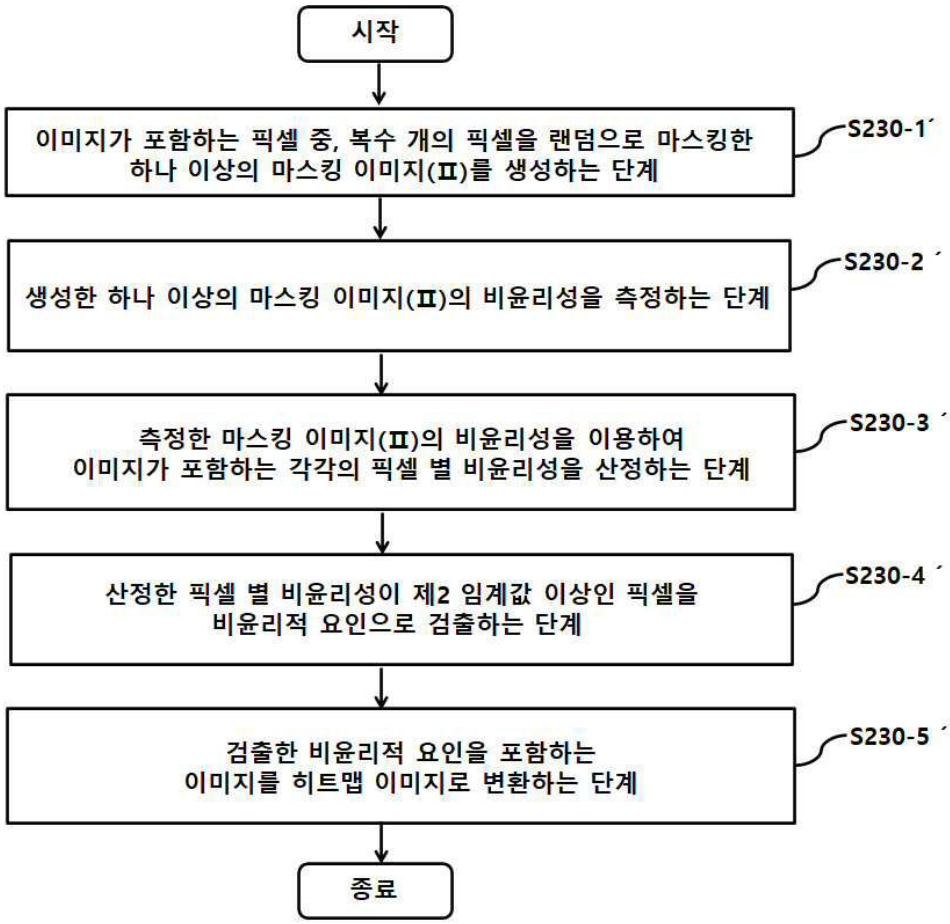
도면5


시작
텍스트를 단어 단위로 분할하는 단계
S230-1
단어 단위로 분할한 텍스트가 포함하는 단어 중
어느 하나를 마스킹하여 하나 이상의 마스킹 텍스트를 생성하는 단계
S230-2
생성한 하나 이상의 마스킹 텍스트에 대응되는
하나 이상의 이미지(I)를 생성하는 단계
S230-3
생성한 마스킹 텍스트에 대응되는
하나 이상의 이미지(I)의 비윤리성을 측정하는 단계
S230-4
측정한 이미지(I)의 비윤리성과 제2 단계에서 측정한 이미지의
비윤리성을 비교하여 비윤리적 요인을 검출하는 단계
S230-5
종료

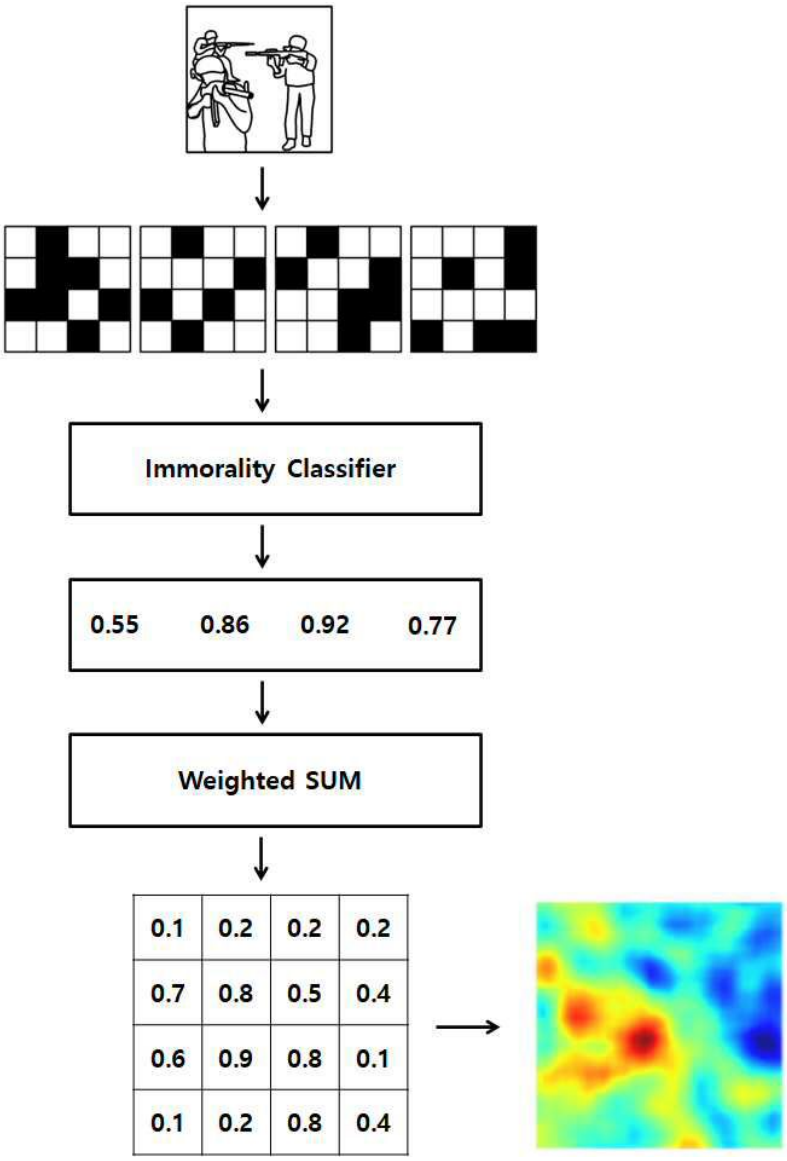
도면6


People shooting a gun at each other
(1) [redacted] shooting a gun at each other
(2) People shooting a [redacted] at each other
(7) People shooting a gun at each [redacted]
Stable Diffusion Model
Immorality Classifier
0.93
0.18
0.89
0.88
비윤리적 요인
People
shooting
a
gun
at
each
other
0.05
-
-
0.7
-
-
0.01

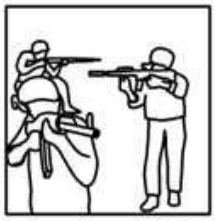
도면7

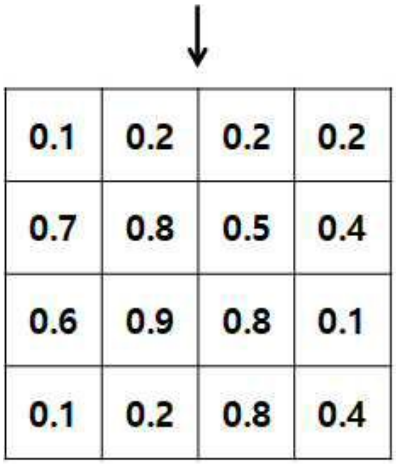

시작
이미지가 포함하는 픽셀 중, 복수 개의 픽셀을 랜덤으로 마스킹한 하나 이상의 마스킹 이미지(Ⅱ)를 생성하는 단계
S230-1´
생성한 하나 이상의 마스킹 이미지(Ⅱ)의 비윤리성을 측정하는 단계
S230-2´
측정한 마스킹 이미지(Ⅱ)의 비윤리성을 이용하여 이미지가 포함하는 각각의 픽셀 별 비윤리성을 산정하는 단계
S230-3´
산정한 픽셀 별 비윤리성이 제2 임계값 이상인 픽셀을 비윤리적 요인으로 검출하는 단계
S230-4´
검출한 비윤리적 요인을 포함하는 이미지를 히트맵 이미지로 변환하는 단계
S230-5´
종료

도면8


Immorality Classifier
0.55 0.86 0.92 0.77
Weighted SUM
0.1 0.2 0.2 0.2
0.7 0.8 0.5 0.4
0.6 0.9 0.8 0.1
0.1 0.2 0.8 0.4

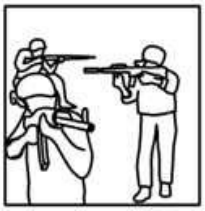
도면9

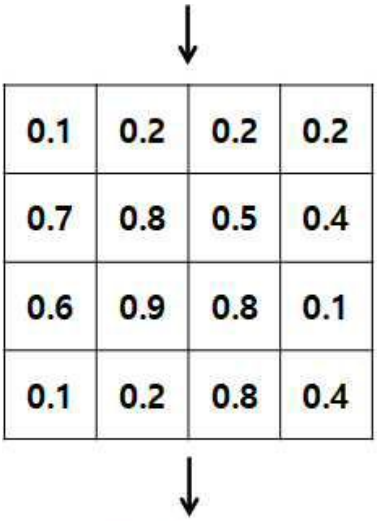




0.1	0.2	0.2	0.2
0.7	0.8	0.5	0.4
0.6	0.9	0.8	0.1
0.1	0.2	0.8	0.4

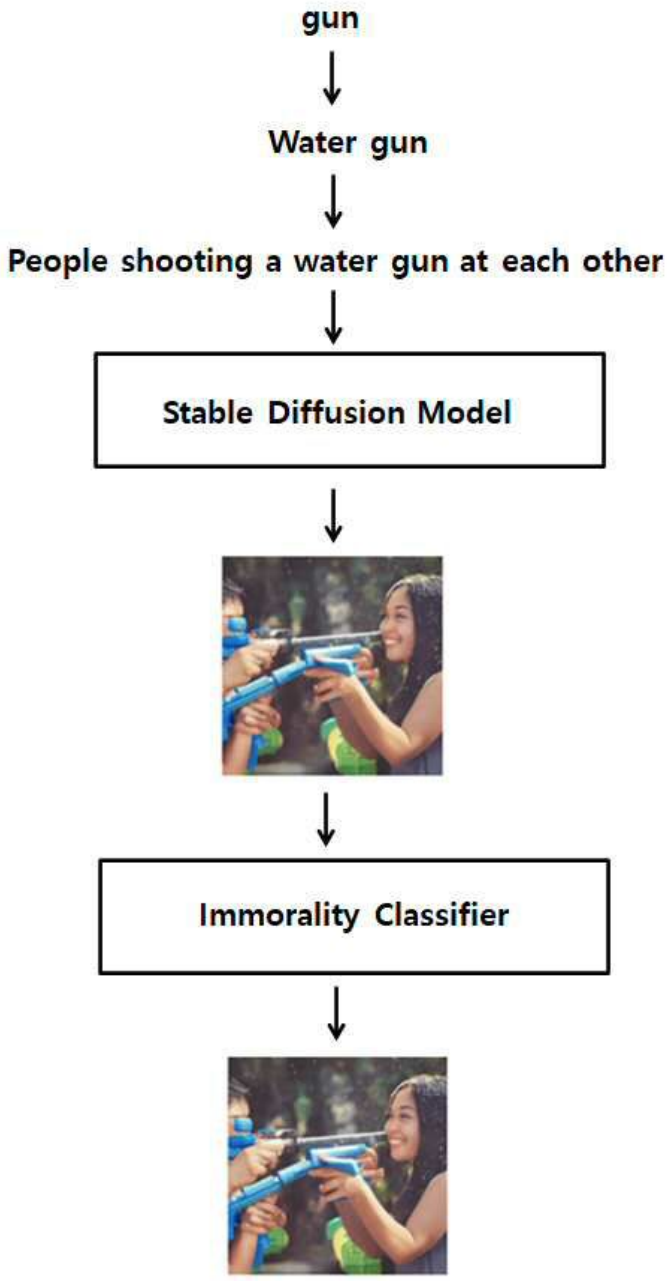
도면10



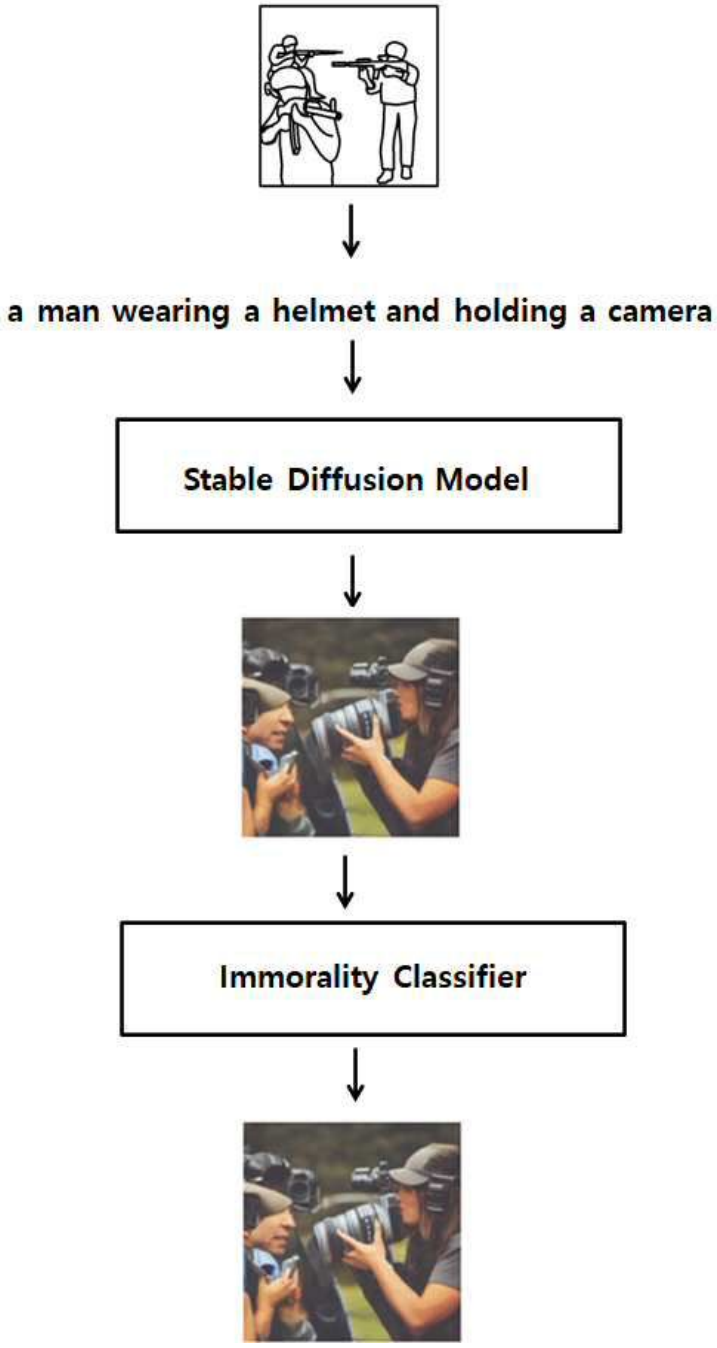


0.1	0.2	0.2	0.2
0.7	0.8	0.5	0.4
0.6	0.9	0.8	0.1
0.1	0.2	0.8	0.4

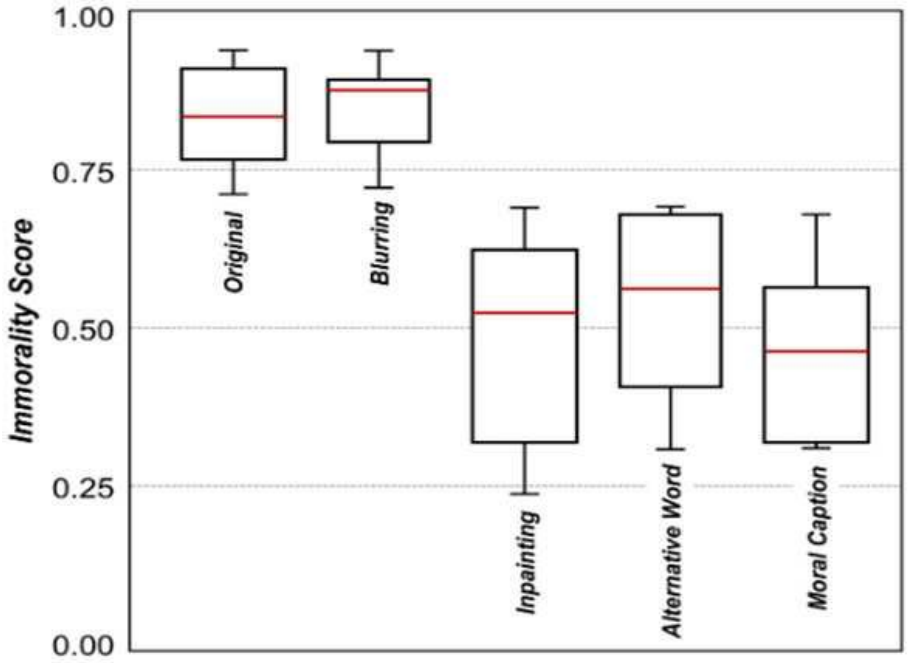
도면11


gun
Water gun
People shooting a water gun at each other
Stable Diffusion Model
Immorality Classifier

도면12


a man wearing a helmet and holding a camera
Stable Diffusion Model
Immorality Classifier

도면13



【심사관 직권보정사항】

【직권보정 1】

【보정항목】 청구범위

【보정세부항목】 청구항 5

【변경전】

제4항에 있어서,

상기 제3-5 단계는,

상기 제3-4 단계에서 측정한 이미지(Ⅰ)의 비윤리성과 상기 제2 단계에서 측정한 이미지의 비윤리성의 차이의 절대값이 가장 큰 이미지(Ⅰ)를 생성하는데 이용한 마스킹 텍스트에서 마스킹된 단어를 비윤릿거 요인으로 검출하는,

비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 방법.

【변경후】

제4항에 있어서,

상기 제3-5 단계는,

상기 제3-4 단계에서 측정한 이미지(Ⅰ)의 비윤리성과 상기 제2 단계에서 측정한 이미지의 비윤리성의 차이의 절대값이 가장 큰 이미지(Ⅰ)를 생성하는데 이용한 마스킹 텍스트에서 마스킹된 단어를 비윤리적 요인으로 검출하는,

비윤리적 이미지를 검출하고 윤리적 이미지로 변환하는 방법.