(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0126247
(43) 공개일자 2024년08월20일

(51) 국제특허분류(Int. Cl.)
G06F 21/55 (2013.01) *G06N 3/094* (2023.01)
G06T 3/40 (2024.01) *G06V 10/764* (2022.01)
(52) CPC특허분류
G06F 21/55 (2013.01)
G06N 3/094 (2023.01)
(21) 출원번호 10-2023-0018803
(22) 출원일자 2023년02월13일
심사청구일자 2023년02월13일

(71) 출원인
한림대학교 산학협력단
강원도 춘천시 한림대학길 1, 한림대학교(옥천동)
(72) 발명자
허종욱
강원도 춘천시 한림대학길 1 공학관 A1301호
이상영
강원도 춘천시 방송길 70(온의동, 온의 롯데캐슬 스카이클래스) 107동 2002호
(74) 대리인
특허법인본

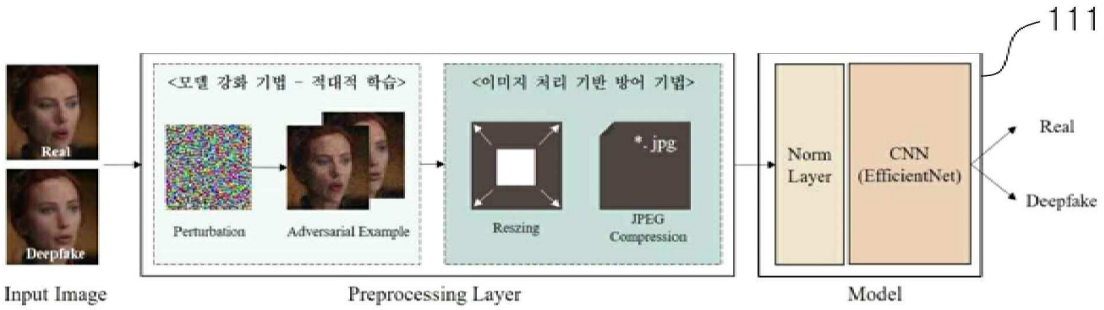
전체 청구항 수 : 총 12 항

(54) 발명의 명칭 **딥페이크 탐지 모델의 적대적 공격 방어를 위한 학습을 수행하는 전자 장치, 전자 장치의 제어 방법, 및 컴퓨터 프로그램**

(57) 요 약

본 발명은, 일반 이미지와 딥페이크(deepfake) 이미지를 분류하기 위한 인공지능 모델을 포함하는 메모리, 및 대상 이미지를 인공지능 모델에 입력하여, 대상 이미지가 일반 이미지인지 딥페이크 이미지인지 여부를 식별하는, 프로세서를 포함하는, 딥페이크 탐지 모델의 적대적 공격 방어를 위한 학습을 수행하는 전자 장치, 전자 장치의 제어 방법, 및 컴퓨터 프로그램에 관한 것이다.

대 표 도 - 도4



(52) CPC특허분류
G06T 3/40 (2024.01)
G06V 10/764 (2023.08)
G06T 2207/20081 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711172885
과제번호	2022R1A4A1033600
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	집단연구지원
연구과제명	자율주행 차량 제어 주체 판별을 위한 디지털 포렌식 원천기술 연구
기 여 율	1/1
과제수행기관명	한림대학교
연구기간	2022.06.01 ~ 2023.02.28

명 세 서

청구범위

청구항 1

전자 장치에 있어서,

일반 이미지와 딥페이크(deepfake) 이미지를 분류하기 위한 인공지능 모델을 포함하는 메모리; 및

대상 이미지를 상기 인공지능 모델에 입력하여, 상기 대상 이미지가 일반 이미지인지 딥페이크 이미지인지 여부를 식별하는, 프로세서;를 포함하고,

상기 인공지능 모델은,

적대적 공격을 포함하는 적대적 예제 이미지를 바탕으로 훈련된 것을 특징으로 하는, 전자 장치.

청구항 2

제1 항에 있어서,

상기 프로세서는,

적어도 하나의 제1 이미지, 및 적어도 하나의 제2 이미지를 포함하는 학습 데이터를 바탕으로, 상기 인공지능 모델을 훈련시키고,

상기 제1 이미지는,

적대적 공격을 미포함하는 이미지이고,

상기 제2 이미지는,

적대적 공격을 포함하는 이미지이고,

상기 학습 데이터는,

일반 이미지 및 딥페이크 이미지 중 적어도 하나를 포함하는 것을 특징으로 하는, 전자 장치.

청구항 3

제2 항에 있어서,

상기 프로세서는,

상기 학습 데이터를 구성하는, 상기 제1 이미지, 및 상기 제2 이미지의 비율을, 각각 40%, 및 60%로 설정하는 것을 특징으로 하는, 전자 장치.

청구항 4

제2 항에 있어서,

상기 프로세서는,

상기 학습 데이터에 포함된 이미지들 각각의 이미지 변환을 수행하는 것을 특징으로 하는, 전자 장치.

청구항 5

제4 항에 있어서,

상기 프로세서는,

상기 학습 데이터에 포함된 이미지들 각각의 크기를 일정 비율로 축소하고,

상기 축소된 이미지들 각각을 이전 크기로 확대한, 제1 집합을 획득하는 것을 특징으로 하는, 전자 장치.

청구항 6

제5 항에 있어서,

상기 프로세서는,

상기 학습 데이터에 포함된 이미지들 각각의 크기를 3배 축소하는 것을 특징으로 하는, 전자 장치.

청구항 7

제5 항에 있어서,

상기 프로세서는,

상기 제1 집합에 포함된 이미지들 각각에 일정 압축 계수를 적용하여 압축한, 제2 집합을 획득하는 것을 특징으로 하는, 전자 장치.

청구항 8

제7 항에 있어서,

상기 프로세서는,

상기 제1 집합에 포함된 이미지들 각각에 적용되는 압축 계수를 15로 설정하는 것을 특징으로 하는, 전자 장치.

청구항 9

제7 항에 있어서,

상기 프로세서는,

상기 제2 집합을 상기 인공지능 모델에 입력하여, 상기 인공지능 모델이 적대적 공격 방어를 수행하도록 학습시키는 것을 특징으로 하는, 전자 장치.

청구항 10

일반 이미지와 딥페이크(deepfake) 이미지를 분류하기 위한 인공지능 모델을 포함하는 메모리; 및

대상 이미지를 상기 인공지능 모델에 입력하여, 상기 대상 이미지가 일반 이미지인지 딥페이크 이미지인지 여부를 식별하는, 프로세서;를 포함하는, 전자 장치의 제어 방법에 있어서,

상기 전자 장치가, 40% 비율의 제1 이미지, 및 60% 비율의 제2 이미지로 구성된 학습 데이터를 획득하는 단계; 및

상기 전자 장치가, 상기 학습 데이터를 바탕으로 상기 인공지능 모델이 일반 이미지와 딥페이크 이미지를 분류하도록 학습시키는 단계;를 포함하고,

상기 제1 이미지는,

적대적 공격을 미포함하는 이미지이고,

상기 제2 이미지는,

적대적 공격을 포함하는 이미지이고,

상기 학습 데이터는,

일반 이미지 및 딥페이크 이미지 중 적어도 하나를 포함하는 것을 특징으로 하는, 전자 장치의 제어 방법.

청구항 11

하드웨어인 컴퓨터와 결합되어, 제10 항의 방법을 수행할 수 있도록 컴퓨터에서 독출가능한 기록매체에 저장된 컴퓨터 프로그램.

청구항 12

일반 이미지와 딥페이크(deepfake) 이미지를 분류하기 위한 인공지능 모델을 포함하는 메모리; 및

대상 이미지를 상기 인공지능 모델에 입력하여, 상기 대상 이미지가 일반 이미지인지 딥페이크 이미지인지 여부를 식별하는, 프로세서;를 포함하는, 전자 장치의 제어 방법에 있어서,

상기 전자 장치가, 적어도 하나의 제1 이미지, 및 적어도 하나의 제2 이미지를 포함하는 학습 데이터를 획득하는 단계;

상기 전자 장치가, 상기 학습 데이터에 포함된 이미지들 각각의 크기를 일정 비율로 축소하는 단계;

상기 전자 장치가, 상기 축소된 이미지들 각각을 이전 크기로 확대한, 제1 집합을 획득하는 단계;

상기 전자 장치가, 상기 제1 집합에 포함된 이미지들 각각에 일정 압축 계수를 적용하여 압축한, 제2 집합을 획득하는 단계; 및

상기 전자 장치가, 상기 제2 집합을 상기 인공지능 모델에 입력하여, 상기 인공지능 모델이 일반 이미지와 딥페이크 이미지를 분류하도록 학습시키는 단계;를 포함하고,

상기 제1 이미지는,

적대적 공격을 미포함하는 이미지이고,

상기 제2 이미지는,

적대적 공격을 포함하는 이미지이고,

상기 학습 데이터는,

일반 이미지 및 딥페이크 이미지 중 적어도 하나를 포함하는 것을 특징으로 하는, 전자 장치의 제어 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 딥페이크(deepfake) 탐지 모델의 적대적 공격 방어를 위한 학습을 수행하는 전자 장치, 전자 장치의 제어 방법, 및 컴퓨터 프로그램에 관한 것으로, 상세하게는, 적대적 공격을 포함하는 딥페이크 이미지 및 일반 이미지와, 적대적 공격을 미포함하는 딥페이크 이미지 및 일반 이미지를 바탕으로 훈련된 인공지능 모델을 통해, 일반 이미지와 딥페이크 이미지를 분류하여, 적대적 공격을 방어하는 기술이다.

배 경 기 술

[0002] 딥페이크(deepfake)로 인한 디지털 범죄는 날로 교묘해지면서 사회적으로 큰 파장을 불러일으키고 있다.

[0003] 예컨대, 소셜 네트워크 서비스에 올렸던 사진이 도용되어 포르노 영상에 합성이 되기도 하고, 유명 인사들의 영상이 조작되어 가짜 뉴스와 같은 결과물이 만들어지기도 한다.

[0004] 이를 가능하게 한 딥페이크 기술은 계속되어 발전하고 있고, 이제 사람의 눈으로는 딥페이크 여부를 판단하기 힘들어졌다.

[0005] 이와 같이, 딥페이크 기술이 발전됨에 따라 딥페이크를 검출해낼 수 있는 딥러닝 기반 모델 또한 계속해서 연구되고 있으나, 딥러닝 기반 모델의 오류를 발생시키는 적대적 공격(adversarial attack)의 등장으로 딥페이크를 탐지하는 모델의 취약성이 증가하고 있고, 이는 매우 치명적인 결과를 초래한다.

[0006] 한편, 상기의 배경기술로서 설명된 사항들은 본 발명의 배경에 대한 이해 증진을 위한 것일 뿐, 이 기술분야에서 통상의 지식을 가진 자에게 이미 알려진 종래기술에 해당함을 인정하는 것으로 받아들여져서는 안 될 것이다.

발명의 내용

해결하려는 과제

[0007] 본 발명이 해결하고자 하는 과제는 적대적 공격을 포함하는 이미지를 적대적 공격을 미포함하는 이미지와 일정 비율로 구성한 학습 데이터를 이미지 처리하는 방법에 따라, 딥페이크 탐지 모델의 적대적 공격 방어를 위한 학

습을 수행하는 전자 장치, 전자 장치의 제어 방법, 및 컴퓨터 프로그램을 제공하는 것이다.

[0008] 구체적으로, 학습 데이터에 적대적 공격이 들어간 이미지(일반 이미지, 딥페이크 이미지 각각에 모두 적용될 수 있다)를 추가하여 학습을 돌리는 기법인 적대적 학습(adversarial training)을 통해 적대적 공격에 대한 저항성을 갖게 하고, 또한 영상처리 기반 방어 기법인 크기 변환(resizing), 및 JPEG 압축을 통해 공격에 대한 섭동 제거하여 적대적 공격에 대한 강인성을 높인 딥페이크 탐지 모델의 적대적 공격 방어를 위한 학습을 수행하는 전자 장치, 전자 장치의 제어 방법, 및 컴퓨터 프로그램을 제공하는 것이다.

[0009] 본 발명이 해결하고자 하는 과제들은 이상에서 언급된 과제로 제한되지 않으며, 언급되지 않은 또 다른 과제들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

과제의 해결 수단

[0010] 상술한 과제를 해결하기 위한 본 발명의 일 면에 따른 전자 장치에 있어서, 일반 이미지와 딥페이크(deepfake) 이미지를 분류하기 위한 인공지능 모델을 포함하는 메모리, 및 대상 이미지를 인공지능 모델에 입력하여, 대상 이미지가 일반 이미지인지 딥페이크 이미지인지 여부를 식별하는, 프로세서를 포함하고, 인공지능 모델은, 적대적 공격을 포함하는 적대적 예제 이미지를 바탕으로 훈련된 것을 특징으로 할 수 있다.

[0011] 추가로, 프로세서는, 적어도 하나의 제1 이미지, 및 적어도 하나의 제2 이미지를 포함하는 학습 데이터를 바탕으로, 인공지능 모델을 훈련시키고, 제1 이미지는, 적대적 공격을 미포함하는 이미지이고, 제2 이미지는, 적대적 공격을 포함하는 이미지이고, 학습 데이터는, 일반 이미지 및 딥페이크 이미지 중 적어도 하나를 포함하는 것을 특징으로 할 수 있다.

[0012] 추가로, 프로세서는, 학습 데이터를 구성하는, 제1 이미지, 및 제2 이미지의 비율을, 각각 40%, 및 60%로 설정하는 것을 특징으로 할 수 있다.

[0013] 추가로, 프로세서는, 학습 데이터에 포함된 이미지들 각각의 이미지 변환을 수행하는 것을 특징으로 할 수 있다.

[0014] 추가로, 프로세서는, 학습 데이터에 포함된 이미지들 각각의 크기를 일정 비율로 축소하고, 축소된 이미지들 각각을 이전 크기로 확대한, 제1 집합을 획득하는 것을 특징으로 할 수 있다.

[0015] 추가로, 프로세서는, 학습 데이터에 포함된 이미지들 각각의 크기를 3배 축소하는 것을 특징으로 할 수 있다.

[0016] 추가로, 프로세서는, 제1 집합에 포함된 이미지들 각각에 일정 압축 계수를 적용하여 압축한, 제2 집합을 획득하는 것을 특징으로 할 수 있다.

[0017] 추가로, 프로세서는, 제1 집합에 포함된 이미지들 각각에 적용되는 압축 계수를 15로 설정하는 것을 특징으로 할 수 있다.

[0018] 추가로, 프로세서는, 제2 집합을 인공지능 모델에 입력하여, 인공지능 모델이 적대적 공격 방어를 수행하도록 학습시키는 것을 특징으로 할 수 있다.

[0019] 한편, 본 발명의 일 면에 따른, 일반 이미지와 딥페이크 이미지를 분류하기 위한 인공지능 모델을 포함하는 메모리, 및 대상 이미지를 인공지능 모델에 입력하여, 대상 이미지가 일반 이미지인지 딥페이크 이미지인지 여부를 식별하는, 프로세서를 포함하는, 전자 장치의 제어 방법에 있어서, 전자 장치가, 40% 비율의 제1 이미지, 및 60% 비율의 제2 이미지로 구성된 학습 데이터를 획득하는 단계, 및 전자 장치가, 학습 데이터를 바탕으로 인공지능 모델이 일반 이미지와 딥페이크 이미지를 분류하도록 학습시키는 단계를 포함하고, 제1 이미지는, 적대적 공격을 미포함하는 이미지이고, 제2 이미지는, 적대적 공격을 포함하는 이미지이고, 학습 데이터는, 일반 이미지 및 딥페이크 이미지 중 적어도 하나를 포함하는 것을 특징으로 한다.

[0020] 한편, 본 발명의 일 면에 따른, 하드웨어인 컴퓨터와 결합되어, 본 발명의 전자 장치의 제어 방법을 수행할 수 있도록 컴퓨터에서 독출가능한 기록매체에 저장된 컴퓨터 프로그램을 더 포함한다.

[0021] 한편, 본 발명의 일 면에 따른, 일반 이미지와 딥페이크 이미지를 분류하기 위한 인공지능 모델을 포함하는 메모리, 및 대상 이미지를 인공지능 모델에 입력하여, 대상 이미지가 일반 이미지인지 딥페이크 이미지인지 여부를 식별하는, 프로세서를 포함하는, 전자 장치의 제어 방법에 있어서, 전자 장치가, 적어도 하나의 제1 이미지, 및 적어도 하나의 제2 이미지를 포함하는 학습 데이터를 획득하는 단계, 전자 장치가, 학습 데이터에 포함된 이미지들 각각의 크기를 일정 비율로 축소하는 단계, 전자 장치가, 축소된 이미지들 각각을 이전 크기로 확대한,

제1 집합을 획득하는 단계, 전자 장치가, 제1 집합에 포함된 이미지들 각각에 일정 압축 계수를 적용하여 압축한, 제2 집합을 획득하는 단계, 및 전자 장치가, 제2 집합을 인공지능 모델에 입력하여, 인공지능 모델이 일반 이미지와 딥페이크 이미지를 분류하도록 학습시키는 단계를 포함하고, 제1 이미지는, 적대적 공격을 미포함하는 이미지이고, 제2 이미지는, 적대적 공격을 포함하는 이미지이고, 학습 데이터는, 일반 이미지 및 딥페이크 이미지 중 적어도 하나를 포함하는 것을 특징으로 한다.

[0022] 본 발명의 기타 구체적인 사항들은 상세한 설명 및 도면들에 포함되어 있다.

발명의 효과

[0023] 본 발명의 딥페이크 탐지 모델의 적대적 공격 방어를 위한 학습을 수행하는 전자 장치, 전자 장치의 제어 방법, 및 컴퓨터 프로그램에 의하면, 학습 데이터의 구성 비율, 및 학습 데이터의 이미지 처리 방식에 의하여, 적대적 공격에도 영향을 받지 않는 강인한(robust) 인공지능 모델을 구현할 수 있다.

[0024] 본 발명의 효과들은 이상에서 언급된 효과로 제한되지 않으며, 언급되지 않은 또 다른 효과들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

[0025] 도 1은 본 개시의 일 실시예에 따른 전자 장치 구성도이다.

도 2는 본 개시의 일 실시예에 따른 기본 흐름도이다.

도 3은 본 개시의 일 실시예에 따른 구체화된 흐름도이다.

도 4는 본 개시의 일 실시예에 따른 인공지능 모델 학습 계통도이다.

도 5는 본 개시의 일 실시예에 따른 적대적 예제 이미지의 비율별 적대적 학습 결과표이다.

도 6은 본 개시의 일 실시예에 따른 이미지 처리 기반 방어 기법 결과표이다.

도 7은 본 개시의 일 실시예에 따른 최종 인공지능 모델 학습 결과표이다.

발명을 실시하기 위한 구체적인 내용

[0026] 본 발명의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 상세하게 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나, 본 발명은 이하에서 개시되는 실시예들에 제한되는 것이 아니라 서로 다른 다양한 형태로 구현될 수 있으며, 단지 본 실시예들은 본 발명의 개시가 완전하도록 하고, 본 발명이 속하는 기술 분야의 통상의 기술자에게 본 발명의 범주를 완전하게 알려주기 위해 제공되는 것이며, 본 발명은 청구항의 범주에 의해 정의될 뿐이다.

[0027] 본 명세서에서 사용된 용어는 실시예들을 설명하기 위한 것이며 본 발명을 제한하고자 하는 것은 아니다. 본 명세서에서, 단수형은 문구에서 특별히 언급하지 않는 한 복수형도 포함한다. 명세서에서 사용되는 "포함한다(comprises)" 및/또는 "포함하는(comprising)"은 언급된 구성요소 외에 하나 이상의 다른 구성요소의 존재 또는 추가를 배제하지 않는다. 명세서 전체에 걸쳐 동일한 도면 부호는 동일한 구성 요소를 지칭하며, "및/또는"은 언급된 구성요소들의 각각 및 하나 이상의 모든 조합을 포함한다. 비록 "제1", "제2" 등이 다양한 구성요소들을 서술하기 위해서 사용되나, 이들 구성요소들은 이들 용어에 의해 제한되지 않음은 물론이다. 이들 용어들은 단지 하나의 구성요소를 다른 구성요소와 구별하기 위하여 사용하는 것이다. 따라서, 이하에서 언급되는 제1 구성요소는 본 발명의 기술적 사상 내에서 제2 구성요소일 수도 있음은 물론이다.

[0028] 다른 정의가 없다면, 본 명세서에서 사용되는 모든 용어(기술 및 과학적 용어를 포함)는 본 발명이 속하는 기술 분야의 통상의 기술자에게 공통적으로 이해될 수 있는 의미로 사용될 수 있을 것이다. 또한, 일반적으로 사용되는 사전에 정의되어 있는 용어들은 명백하게 특별히 정의되어 있지 않는 한 이상적으로 또는 과도하게 해석되지 않는다.

[0029] 명세서에서 사용되는 "부" 또는 “모듈”이라는 용어는 소프트웨어, FPGA 또는 ASIC과 같은 하드웨어 구성요소를 의미하며, "부" 또는 “모듈”은 어떤 역할들을 수행한다. 그렇지만 "부" 또는 “모듈”은 소프트웨어 또는 하드웨어에 한정되는 의미는 아니다. "부" 또는 “모듈”은 어드레싱할 수 있는 저장 매체에 있도록 구성될 수도 있고 하나 또는 그 이상의 프로세서들을 재생시키도록 구성될 수도 있다. 따라서, 일 예로서 "부" 또는 “모듈”은 소프트웨어 구성요소들, 객체지향 소프트웨어 구성요소들, 클래스 구성요소들 및 태스크 구성요소들과

같은 구성요소들과, 프로세스들, 함수들, 속성들, 프로시저들, 서브루틴들, 프로그램 코드의 세그먼트들, 드라이버들, 펌웨어, 마이크로 코드, 회로, 데이터, 데이터베이스, 데이터 구조들, 테이블들, 어레이들 및 변수들을 포함한다. 구성요소들과 "부" 또는 “모듈”들 안에서 제공되는 기능은 더 작은 수의 구성요소들 및 "부" 또는 “모듈”들로 결합되거나 추가적인 구성요소들과 "부" 또는 “모듈”들로 더 분리될 수 있다.

[0030] 공간적으로 상대적인 용어인 "아래(below)", "아래(beneath)", "하부(lower)", "위(above)", "상부(upper)" 등은 도면에 도시되어 있는 바와 같이 하나의 구성요소와 다른 구성요소들과의 상관관계를 용이하게 기술하기 위해 사용될 수 있다. 공간적으로 상대적인 용어는 도면에 도시되어 있는 방향에 더하여 사용시 또는 동작시 구성요소들의 서로 다른 방향을 포함하는 용어로 이해되어야 한다. 예를 들어, 도면에 도시되어 있는 구성요소를 뒤집을 경우, 다른 구성요소의 "아래(below)"또는 "아래(beneath)"로 기술된 구성요소는 다른 구성요소의 "위(above)"에 놓일 수 있다. 따라서, 예시적인 용어인 "아래"는 아래와 위의 방향을 모두 포함할 수 있다. 구성요소는 다른 방향으로도 배향될 수 있으며, 이에 따라 공간적으로 상대적인 용어들은 배향에 따라 해석될 수 있다.

[0031] 본 명세서에서, 컴퓨터는 적어도 하나의 프로세서를 포함하는 모든 종류의 하드웨어 장치를 의미하는 것이고, 실시 예에 따라 해당 하드웨어 장치에서 동작하는 소프트웨어적 구성도 포괄하는 의미로서 이해될 수 있다. 예를 들어, 컴퓨터는 스마트폰, 태블릿 PC, 데스크톱, 노트북 및 각 장치에서 구동되는 사용자 클라이언트 및 애플리케이션을 모두 포함하는 의미로서 이해될 수 있으며, 또한 이에 제한되는 것은 아니다.

[0032] 이하, 첨부된 도면을 참조하여 본 발명의 실시예를 상세하게 설명한다.

[0033] 도 1은 본 개시의 일 실시예에 따른 전자 장치(100) 구성도이다.

[0034] 도 1에 도시된 바와 같이, 전자 장치(100)는, 일반 이미지와 딥페이크 이미지를 분류하기 위한 인공지능 모델(111)을 포함하는 메모리(110), 및 대상 이미지를 인공지능 모델(111)에 입력하여, 대상 이미지가 일반 이미지인지 딥페이크 이미지인지 여부를 식별하는, 프로세서(120)를 포함한다.

[0035] 인공지능 모델(111)은, 전자 장치(100)에 의해, 적대적 공격을 포함하는 딥페이크 이미지에 해당하는 적대적 예제 이미지를 바탕으로 훈련된 모델일 수 있다. 구체적으로, 인공지능 모델(111)은 적대적 공격(adversarial attack)을 미포함하는 이미지인, 제1 이미지 및 적대적 공격을 포함하는 이미지인, 제2 이미지를 바탕으로 훈련될 수 있으며, 입력된 이미지가 딥페이크 이미지인지 일반 이미지인지 여부를 구분하도록 훈련될 수 있다. 이때, 제1 이미지와 제2 이미지 간의 비율은 4 대 6일 수 있으나, 이에 한정되지 않는다. 이는, 후술할 도 5와 관련된다.

[0036] 여기서, 일반 이미지는, 딥페이크 이미지가 아닌 일반적인 이미지를 의미한다. 즉, 조작/위조되지 않은 이미지를 의미할 수 있다.

[0037] 이때, 제1 이미지 및 제2 이미지를 포함하여, 인공지능 모델(111)을 훈련시키는데 사용되는, 훈련 데이터는, 일반 이미지 및 딥페이크 이미지 중 적어도 하나를 포함할 수 있다.

[0038] 구체적으로, 학습 데이터는, 제1 이미지인 일반 이미지, 제2 이미지인 일반 이미지, 제1 이미지인 딥페이크 이미지, 및 제2 이미지인 딥페이크 이미지 중 적어도 하나를 포함할 수 있다.

[0039] 전자 장치(100)는, 서버, 스마트폰(smartphone), 태블릿 PC(tablet personal computer), 이동 전화기(mobile phone), 영상 전화기, 랩탑 PC(laptop PC), 넷북 컴퓨터(netbook computer), 노트북 컴퓨터(notebook computer), 웨어러블 장치(wearable device) 중 적어도 하나를 포함할 수 있다.

[0040] 한편, 메모리(110)는 전자 장치(100)의 동작에 필요한 각종 프로그램 및 데이터를 저장할 수 있다. 메모리(110)는 비휘발성 메모리(110), 휘발성 메모리(110), 플래시메모리(110)(flash-memory), 하드디스크 드라이브(HDD) 또는 솔리드 스테이트 드라이브(SSD) 등으로 구현될 수 있다.

[0041] 통신부는 외부 장치와 통신을 수행할 수 있다. 특히, 통신부는 와이파이 칩, 블루투스 칩, 무선 통신 칩, NFC칩, 저전력 블루투스 칩(BLE 칩) 등과 같은 다양한 통신 칩을 포함할 수 있다. 이때, 와이파이 칩, 블루투스 칩, NFC 칩은 각각 LAN 방식, WiFi 방식, 블루투스 방식, NFC 방식으로 통신을 수행한다. 와이파이 칩이나 블루투스칩을 이용하는 경우에는 SSID 및 세션 키 등과 같은 각종 연결 정보를 먼저 송수신 하여, 이를 이용하여 통신 연결한 후 각종 정보들을 송수신할 수 있다. 무선 통신칩은 IEEE, 지그비, 3G(3rd Generation), 3GPP(3rd Generation Partnership Project), LTE(Long Term Evolution), 5G(5th Generation) 등과 같은 다양한 통신 규

격에 따라 통신을 수행하는 칩을 의미한다.

[0042] 프로세서(120)는 메모리(110)에 저장된 각종 프로그램을 이용하여 전자 장치(100)의 전반적인 동작을 제어할 수 있다. 프로세서(120)는 RAM, ROM, 그래픽 처리부, 메인 CPU, 제1 내지 n 인터페이스 및 버스로 구성될 수 있다. 이때, RAM, ROM, 그래픽 처리부, 메인 CPU, 제1 내지 n 인터페이스 등은 버스를 통해 서로 연결될 수 있다.

[0043] RAM은 O/S 및 어플리케이션 프로그램을 저장한다. 구체적으로, 전자 장치(100)가 부팅되면 O/S가 RAM에 저장되고, 사용자가 선택한 각종 어플리케이션 데이터가 RAM에 저장될 수 있다.

[0044] ROM에는 시스템 부팅을 위한 명령어 세트 등이 저장된다. 턴 온 명령이 입력되어 전원이 공급되면, 메인 CPU는 ROM에 저장된 명령어에 따라 메모리(110)에 저장된 O/S를 RAM에 복사하고, O/S를 실행시켜 시스템을 부팅시킨다. 부팅이 완료되면, 메인 CPU는 메모리(110)에 저장된 각종 어플리케이션 프로그램을 RAM에 복사하고, RAM에 복사된 어플리케이션 프로그램을 실행시켜 각종 동작을 수행한다.

[0045] 그래픽 처리부는 연산부(미도시) 및 렌더링부(미도시)를 이용하여 아이템, 이미지, 텍스트 등과 같은 다양한 객체를 포함하는 화면을 생성한다. 여기서, 연산부는 입력부로부터 수신된 제어 명령을 이용하여 화면의 레이아웃에 따라 각 객체들이 표시될 좌표값, 형태, 크기, 컬러 등과 같은 속성값을 연산하는 구성일 수 있다. 그리고, 렌더링부는 연산부에서 연산한 속성값에 기초하여 객체를 포함하는 다양한 레이아웃의 화면을 생성하는 구성이일 수 있다. 이러한 렌더링부에서 생성된 화면은 디스플레이의 디스플레이 영역 내에 표시될 수 있다.

[0046] 메인 CPU는 메모리(110)에 액세스하여, 메모리(110)에 저장된 OS를 이용하여 부팅을 수행한다. 그리고, 메인 CPU는 메모리(110)에 저장된 각종 프로그램, 컨텐츠, 데이터 등을 이용하여 다양한 동작을 수행한다.

[0047] 제1 내지 n 인터페이스는 상술한 각종 구성요소들과 연결된다. 제1 내지 n 인터페이스 중 하나는 네트워크를 통해 외부 장치와 연결되는 네트워크 인터페이스가 될 수도 있다.

[0048] 한편, 나아가, 프로세서(120)는 인공지능 모델(111)을 제어할 수 있다. 이 경우, 프로세서는 인공지능 모델을 제어하기 위한 그래픽 전용 프로세서(예: GPU)를 포함할 수 있음은 물론이다.

[0049] 한편, 본 발명에 따른 인공지능 모델(111)은 지도학습(supervised learning) 또는 비지도학습(unsupervised learning)기반의 모델일 수 있다. 나아가, 본 발명에 따른 인공지능 모델(111)은 SVM(support vector machine), Decision tree, neural network 등 및 이들이 응용된 방법론을 포함할 수 있다.

[0050] 대표적으로, 본 발명에 따른 인공지능 모델(111)은 학습데이터를 입력하여 학습된 합성곱 신경망(Convolutional deep Neural Networks, CNN) 기반의 인공지능 모델일 수 있다. 다만, 이에 한정되는 것은 아니며, 다양한 인공지능 모델이 본 발명에 적용될 수 있음은 물론이다. 예컨대, DNN(Deep Neural Network), RNN(Recurrent Neural Network), BRDNN(Bidirectional Recurrent Deep Neural Network), RF(Random Forest)와 같은 모델이 인공지능 모델로서 사용될 수 있으나, 이에 한정되지 않는다.

[0051] 이때, 합성곱 신경망(Convolutional deep Neural Networks, CNN)은 최소한의 전처리(preprocess)를 사용하도록 설계된 다계층 퍼셉트론(multilayer perceptrons)의 한 종류이다. 합성곱 신경망은 하나 또는 여러개의 합성곱 계층(convolutional layer)과 그 위에 올려진 일반적인 인공신경망 계층들로 이루어져 있으며, 가중치와 통합 계층(pooling layer)들을 추가로 활용한다. 이러한 구조 덕분에 합성곱 신경망은 2차원 구조의 입력 데이터를 충분히 활용할 수 있다. 또한, 합성곱 신경망은 표준 역전달을 통해 훈련될 수 있다. 합성곱 신경망은 다른 피드포워드 인공신경망 기법들보다 쉽게 훈련되는 편이고 적은 수의 매개변수를 사용한다는 이점이 있다.

[0052] 또한, 심층 신경망(Deep Neural Networks, DNN)은 입력 계층(input layer)과 출력 계층(output layer) 사이에 복수개의 은닉 계층(hidden layer)들로 이뤄진 인공신경망(Artificial Neural Network, ANN)이다.

[0053] 이때, 심층 신경망의 구조는 퍼셉트론(perceptron)으로 구성될 수 있다. 퍼셉트론은 여러 개의 입력 값(input)과 하나의 프로세서(prosessor), 하나의 출력 값으로 구성된다. 프로세서는 여러 개의 입력 값에 각각 가중치를 곱한 후, 가중치가 곱해진 입력 값들을 모두 합한다. 그 다음 프로세서는 합해진 값을 활성화함수에 대입하여 하나의 출력 값을 출력한다. 만약 활성화함수의 출력 값으로 특정한 값이 나오기를 원하는 경우, 각 입력 값에 곱해지는 가중치를 수정하고, 수정된 가중치를 이용하여 출력 값을 다시 계산할 수 있다. 이때, 각각의 퍼셉트론은 서로 다른 활성화함수를 사용할 수 있다. 또한 각각의 퍼셉트론은 이전 계층에서 전달된 출력들을 입력으로 받아들인 다음, 활성화 함수를 이용해서 출력을 구한다. 구해진 출력은 다음 계층의 입력으로 전달된다. 상술한 바와 같은 과정을 거치면 최종적으로 몇 개의 출력 값을 얻을 수 있다.

[0054] 순환 신경망(Reccurent Neural Network, RNN)은 인공신경망을 구성하는 유닛 사이의 연결이 Directed cycle을 구성하는 신경망을 말한다. 순환 신경망은 앞먹임 신경망과 달리, 임의의 입력을 처리하기 위해 신경망 내부의 메모리를 활용할 수 있다.

[0055] 심층 신뢰 신경망(Deep Belief Networks, DBN)이란 기계학습에서 사용되는 그래프 생성 모형(generative graphical model)으로, 딥 러닝에서는 잠재변수(latent variable)의 다중계층으로 이루어진 심층 신경망을 의미한다. 계층 간에는 연결이 있지만 계층 내의 유닛 간에는 연결이 없다는 특징이 있다.

[0056] 심층 신뢰 신경망은 생성 모형이라는 특성상 선행학습에 사용될 수 있고, 선행학습을 통해 초기 가중치를 학습한 후 역전파 혹은 다른 판별 알고리즘을 통해 가중치의 미조정을 할 수 있다. 이러한 특성은 훈련용 데이터가 적을 때 굉장히 유용한데, 이는 훈련용 데이터가 적을수록 가중치의 초기값이 결과적인 모델에 끼치는 영향이 세지기 때문이다. 선행학습된 가중치 초기값은 임의로 설정된 가중치 초기값에 비해 최적의 가중치에 가깝게 되고 이는 미조정 단계의 성능과 속도향상을 가능케 한다.

[0057] RF는, 분류, 회기 분석 등에 사용되는 앙상블 학습 방법의 일종으로, 훈련 과정에서 구성한 다수의 결정 트리로부터 분류 또는 평균 예측치(회귀 분석)를 출력함으로써 동작하며, 가장 큰 특징은 랜덤성(randomness)에 의해 트리들이 서로 조금씩 다른 특성을 갖는다는 점이다. 이 특성은 각 트리들의 예측(prediction)들이 비상관화(decorrelation) 되게하며, 결과적으로 일반화(generalization) 성능을 향상시킨다. 또한, 랜덤화(randomization)는 포레스트가 노이즈가 포함된 데이터에 대해서도 강인하게 만들어 준다. 랜덤화는 각 트리들의 훈련 과정에서 진행되며, 랜덤 학습 데이터 추출 방법을 이용한 앙상블 학습법인 배깅(bagging)과 랜덤 노드 최적화(randomized node optimization)가 자주 사용된다. 이 두 가지 방법은 서로 동시에 사용되어 랜덤화 특성을 더욱 증진시킬 수 있다.

[0058] 상술한 인공지능 및 그 학습방법에 관한 내용은 예시를 위하여 서술된 것이며, 상술한 실시 예들에서 이용되는 인공지능 및 그 학습방법은 제한되지 않는다. 예를 들어, 당 업계의 통상의 기술자가 동일한 과제해결을 위하여 적용할 수 있는 모든 종류의 인공지능 기술 및 그 학습방법이 개시된 실시 예에 따른 시스템을 구현하는 데 활용될 수 있다.

[0059] 한편, 프로세서(120)는 하나 이상의 코어(core, 미도시) 및 그래픽 처리부(미도시) 및/또는 다른 구성 요소와 신호를 송수신하는 연결 통로(예를 들어, 버스(bus) 등)를 포함할 수 있다.

[0060] 일 실시예에 따른 프로세서(120)는 메모리(110)에 저장된 하나 이상의 인스트럭션을 실행함으로써, 본 발명과 관련하여 설명된 방법을 수행한다.

[0061] 예를 들어, 프로세서(120)는 메모리(110)에 저장된 하나 이상의 인스트럭션을 실행함으로써 신규 학습용 데이터를 획득하고, 학습된 모델을 이용하여, 상기 획득된 신규 학습용 데이터에 대한 테스트를 수행하고, 상기 테스트 결과, 라벨링된 정보가 소정의 제1 기준값 이상의 정확도로 획득되는 제1 학습용 데이터를 추출하고, 상기 추출된 제1 학습용 데이터를 상기 신규 학습용 데이터로부터 삭제하고, 상기 추출된 학습용 데이터가 삭제된 상기 신규 학습용 데이터를 이용하여 상기 학습된 모델을 다시 학습시킬 수 있다.

[0062] 한편, 프로세서(120)는 프로세서(120) 내부에서 처리되는 신호(또는, 데이터)를 일시적 및/또는 영구적으로 저장하는 램(RAM: Random Access Memory, 미도시) 및 롬(ROM: Read-Only Memory, 미도시)을 더 포함할 수 있다. 또한, 프로세서(120)는 그래픽 처리부, 램 및 롬 중 적어도 하나를 포함하는 시스템온칩(SoC: system on chip) 형태로 구현될 수 있다.

[0063] 메모리(110)에는 프로세서(120)의 처리 및 제어를 위한 프로그램들(하나 이상의 인스트럭션들)을 저장할 수 있다. 메모리(110)에 저장된 프로그램들은 기능에 따라 복수 개의 모듈들로 구분될 수 있다.

[0064] 도 2는 본 개시의 일 실시예에 따른 기본 흐름도이다.

[0065] 도 2에 도시된 바와 같이, 전자 장치(100)는, 적어도 하나의 제1 이미지, 및 적어도 하나의 제2 이미지를 포함하는 학습 데이터를 획득(S210)하여, 학습 데이터를 바탕으로, 인공지능 모델(111)을 훈련(S220)시킬 수 있다.

[0066] 도 3은 본 개시의 일 실시예에 따른 구체화된 흐름도이고, 도 4는 본 개시의 일 실시예에 따른 인공지능 모델 학습 계통도이다.

[0067] 도 3 및 도 4에 도시된 바와 같이, 전자 장치(100)는 단계 S210을 수행하여 획득된 학습 데이터에 포함된 이미지들 각각의 이미지 변환을 수행할 수 있다.

[0068] 구체적으로, 전자 장치(100)는, 학습 데이터에 포함된 이미지들 각각의 크기를 일정 비율로 축소(S310)하고, 축소된 이미지들 각각을 이전 크기로 확대한, 제1 집합을 획득(S320)할 수 있다.

[0069] 이어서, 전자 장치(100)는, 제1 집합에 포함된 이미지들 각각에 일정 압축 계수를 적용하여 압축한, 제2 집합을 획득(S330)함에 따라, 단계 S220을 수행함에 있어서, 제2 집합을 인공지능 모델(111)에 입력하여, 인공지능 모델(111)이 적대적 공격 방어를 수행하도록 학습시킬 수 있다.

[0070] 집합은, 복수의 이미지를 포함하는 데이터이다.

[0071] 이에 따라, 전자 장치(100)는, 딥러닝 기반 딥페이크 탐지 모델에 치명적 오류를 범하게 하는 적대적 공격에 대하여 크기 변환 및 압축의 2가지 기법을 통해 강인한 인공지능 모델(111)을 구축할 수 있다.

[0072] 한편, 전자 장치(100)가, 단계 S310 내지 S330의 이미지 전처리에서 정규화를 적용할 경우, 픽셀 값이 [-1, 1]로 출력되는데, 적대적 공격 모듈인 torchattacks의 입력값은 [0, 1]의 픽셀 값으로 맞춰 구현되어 있기 때문에, 적대적 공격을 적용한 뒤 정규화를 진행해야만 한다.

[0073] 이를 위한, 실시예로, 전자 장치(100)는, 도 4에 도시된 바와 같이, 인공지능 모델(111)에 정규화 역할을 수행하는 레이어(layer)를 추가할 수 있다.

[0074] 구체적으로, 단계 S330을 수행함에 있어서, 전자 장치(100)는, 제2 집합에 포함된 이미지들 각각의 픽셀 값을, 기 설정된 기준 값(ex. [0, 1])으로 정규화한, 제3 집합을 획득할 수 있다.

[0075] 이후, 단계 S220을 수행함에 있어서, 전자 장치(100)는, 제3 집합을 인공지능 모델(111)에 입력하여, 인공지능 모델(111)을 학습시킬 수 있다.

[0076] 실시예로, 단계 S210을 수행함에 있어서, 전자 장치(100)는, 적어도 하나의 일반 이미지와 적어도 하나의 딥페이크 이미지로 학습 데이터를 구성하되, 일반 이미지, 및 딥페이크 이미지 각각은, 적대적 요소가 있는 적대적 예제 이미지, 및 적대적 요소가 없는 비적대적 예제 이미지를 포함할 수 있다.

[0077] 즉, 학습 데이터는, 제1 이미지인 일반 이미지, 제2 이미지인 일반 이미지, 제1 이미지인 딥페이크 이미지, 및 제2 이미지인 딥페이크 이미지 중 적어도 하나를 포함할 수 있다.

[0078] 한편, 전자 장치(100)는, 일반 이미지와 딥페이크 이미지로 구성된 비교 학습 데이터를 바탕으로 일반 이미지와 딥페이크 이미지를 구분하도록 학습된 종래의 인공지능 모델인, 비교 모델을 메모리(110)에 더 포함할 수 있다. 여기서, 비교 학습 데이터에 포함된 이미지들은, 적대적 공격을 포함하지 않을 수 있다.

[0079] 이때, 전자 장치(100)는 학습 데이터 중 적대적 예제 이미지를 인공지능 모델(111) 및 비교 모델에 각각 입력하여, 적대적 예제 이미지의 차별적 특징과 관련된 데이터를 획득할 수 있다. 구체적으로, 적대적 예제 이미지 중 비교 모델에 의해 적발될 수 없는 이미지의 차별적 특징이 식별될 수 있다.

[0080] 관련하여, 일 실시예로, 전자 장치(100)는, 학습 데이터를 구성하는 이미지들을 인공지능 모델(111)과 비교 모델에 각각 입력하여, 인공지능 모델(111)의 출력 및 비교 모델의 출력을 각각 획득할 수 있다.

[0081] 여기서, 전자 장치(100)는 인공지능 모델(111)의 출력과 비교 모델의 출력이 다르게 나타난 적어도 하나의 이미지를 식별할 수 있다. 구체적인 예로, 전자 장치(100)는 인공지능 모델(111)에 의해서는 딥페이크 이미지로 식별되었으나 비교 모델에 의해서는 일반 이미지로 식별된 적대적 예제 이미지(: 이하 타겟 이미지로 지칭함)를 식별할 수 있다.

[0082] 이때, 전자 장치(100)는, 타겟 이미지들의 공통된 특징 데이터를 획득할 수 있다. 이를 위해, 전자 장치(100)는 타겟 이미지들과 다른 이미지들 간의 차이를 파라미터 별로 분석하여 타겟 이미지의 특징 데이터를 획득할 수 있다.

[0083] 이때, 특징 데이터는 이미지의 특징과 관련된 하나 이상의 파라미터의 값에 해당할 수 있다. 예를 들어, 파라미터는, sharpness, contrast, resolution, gradation 등에 해당할 수 있으나, 이에 한정되지 않는다.

[0084] 전자 장치(100)는, 타겟 이미지 각각의 특징 데이터를 바탕으로, 파라미터 별로 타겟 이미지들의 수치 범위를 식별하여 저장할 수 있다. 이후, 적어도 하나의 이미지에 대한 파라미터가 분석된 결과, 해당 이미지가 타겟 이미지에 해당할지 여부가 판별될 수 있다.

[0085] 관련하여, 일 실시 예로, 전자 장치(100)는 입력되는 이미지의 파라미터 별 값에 따라 인공지능 모델(111) 및

비교 모델 중 적어도 하나를 선택적으로 활용할 수 있게 된다.

[0086] 구체적으로, 전자 장치(100)는 입력되는 이미지의 파라미터 별 값을 상술한 특징 데이터와 비교하여, 입력된 이미지가 타겟 이미지(: 적대적 예제 이미지)에 해당하는지 여부를 식별할 수 있다.

[0087] 여기서, 입력된 이미지가 타겟 이미지에 해당하는 경우, 전자 장치(100)는 입력된 이미지를 인공지능 모델(111)에 입력하여 딥페이크 이미지인지 여부를 식별할 수 있다.

[0088] 반면, 입력된 이미지가 타겟 이미지에 해당하지 않는 경우, 전자 장치(100)는 입력된 이미지를 비교 모델에 입력하여 딥페이크 이미지인지 여부를 식별할 수 있다.

[0089] 그 결과, 적대적 공격을 포함하는, 즉 제2 이미지인 딥페이크 이미지일 확률이 높은 이미지에 대해서는 본 개시에 따른 인공지능 모델(111)이 활용되고, 일반 이미지이거나, 적대적 공격을 포함하지 않는 딥페이크 이미지(제1 이미지인 딥페이크 이미지)일 확률이 높은 이미지에 대해서는 종래 기술에 따른 상술한 비교 모델이 이용될 수 있다. 상술한 실시 예에 따라, 인공지능 모델(111)의 구동을 위한 연산량이 비교 모델보다 더 높을 수 있으므로, 이렇듯 이미지의 종류를 구분하여 인공지능 모델을 선택적으로 활용함으로써 연산량을 최소화하면서도 판단의 정확도는 유지할 수 있다.

[0090] 한편, 도 5는 본 개시의 일 실시예에 따른 적대적 예제 이미지의 비율별 적대적 학습 결과표이고, 도 6은 본 개시의 일 실시예에 따른 이미지 처리 기반 방어 기법 결과표이고, 도 7은 본 개시의 일 실시예에 따른 최종 인공지능 모델 학습 결과표이다.

[0091] 도 5에 도시된 바와 같이, 제1 이미지(original sample)와 제2 이미지(adversarial example)의 비율이 4 대 6으로 설정했을 때, 가장 높은 성능이 제1 이미지와 제2 이미지 각각에 대해 고르게 발생한 것으로 나타났다.

[0092] 따라서, 단계 S210을 수행함에 있어서, 전자 장치(100)는, 학습 데이터를 구성하는, 제1 이미지, 및 제2 이미지의 비율을, 각각 40%, 및 60%로 설정할 수 있다.

[0093] 도 6에 도시된 바와 같이, 이미지 크기 변환(Resize)과 JEPG 압축에 있어서, 이미지 크기 변환에서는 이미지를 3배 압축하여, 압축된 이미지를 3배 확대하여 원래 크기로 변환하여 입력값으로 사용하였을 때가 가장 높은 공격 방어율을 보이는 것으로 나타났다.

[0094] 한편, 종래 기술에 의하면, JEPG 압축의 경우, 압축계수를 25미만으로 설정하였을 때, 높은 방어율을 확보할 수 있다는 주장(Dong, Xiaoyi, et al. "Robust superpixel-guided attentional adversarial attack." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020)이 있어왔으나, 도 6에 도시된 바와 같이, 압축계수를 15로 한정하였을 때, 가장 좋은 결과를 나타낸 것으로 나타났다.

[0095] 또한, 선형적 변환인 크기 조정보다, 비선형적 변환인 압축을 사용했을 때, 공격 방어율이 더 높은 것을, 도 5 및 도 6에서 확인할 수 있다.

[0096] 따라서, 단계 S310 내지 S330을 수행함에 있어서, 전자 장치(100)는, 학습 데이터에 포함된 이미지들 각각의 크기를 3배 축소하고, 제1 이미지 집합에 포함된 이미지들 각각에 적용되는 압축 계수를 15로 설정할 수 있다.

[0097] 또한, 도 7에 도시된 바와 같이, FGSM 뿐만 아니라, BIM, PGD와 같은 기법을 학습데이터에 모두 적용하였을 때의 결과가 단일 방법만 적용했을 때 보다 높은 성능을 보이는 것으로 나타났다.

[0098] 이에 따라, 전자 장치(100)는, 학습 데이터를 이미지 처리함에 있어서, 크기 조정 및 압축을 모두 적용함으로써 인공지능 모델의 딥페이크에 대한 방어 성능을 높일 수 있다.

[0099] 한편, 일 실시예로, 전자 장치(100)의 제어 방법에 의하면, 전자 장치(100)는, 40% 비율의 제1 이미지, 및 60% 비율의 제2 이미지로 구성된 학습 데이터를 획득하고, 학습 데이터를 바탕으로 인공지능 모델(111)이 일반 이미지와 딥페이크 이미지를 분류하도록 학습시킬 수 있다.

[0100] 한편, 다른 실시예로, 전자 장치(100)의 제어 방법에 의하면, 전자 장치(100)는, 적어도 하나의 제1 이미지, 및 적어도 하나의 제2 이미지를 포함하는 학습 데이터를 획득할 수 있다.

[0101] 이에 따라, 전자 장치(100)는, 학습 데이터에 포함된 이미지들 각각의 크기를 일정 비율로 축소하고, 축소된 이미지들 각각을 이전 크기로 확대한, 제1 집합을 획득하여, 제1 집합에 포함된 이미지들 각각에 일정 압축 계수를 적용하여 압축한, 제2 집합을 획득할 수 있다.

[0102] 전자 장치(100)는, 제2 집합을 인공지능 모델(111)에 입력하여, 인공지능 모델(111)이 일반 이미지와 딥페이크

이미지를 분류하도록 학습시킬 수 있다.

[0103] 한편, 본 발명의 일 면에 따른, 하드웨어인 컴퓨터와 결합되어, 상술한 본 발명의 전자 장치(100)의 제어 방법을 수행할 수 있도록 컴퓨터에서 독출가능한 기록매체에 저장된 컴퓨터 프로그램을 더 포함한다.

[0104] 본 발명의 실시예와 관련하여 설명된 방법 또는 알고리즘의 단계들은 하드웨어로 직접 구현되거나, 하드웨어에 의해 실행되는 소프트웨어 모듈로 구현되거나, 또는 이들의 결합에 의해 구현될 수 있다. 소프트웨어 모듈은 RAM(Random Access Memory), ROM(Read Only Memory), EPROM(Erasable Programmable ROM), EEPROM(Electrically Erasable Programmable ROM), 플래시 메모리(Flash Memory), 하드 디스크, 착탈형 디스크, CD-ROM, 또는 본 발명이 속하는 기술 분야에서 잘 알려진 임의의 형태의 컴퓨터 판독가능 기록매체에 상주할 수도 있다.

[0105] 또한, 본 발명의 서로 다른 실시예들은 상호 보완되거나 결합될 수 있다.

[0106] 본 발명의 구성 요소들은 하드웨어인 컴퓨터와 결합되어 실행되기 위해 프로그램(또는 애플리케이션)으로 구현되어 매체에 저장될 수 있다. 본 발명의 구성 요소들은 소프트웨어 프로그래밍 또는 소프트웨어 요소들로 실행될 수 있으며, 이와 유사하게, 실시 예는 데이터 구조, 프로세스들, 루틴들 또는 다른 프로그래밍 구성들의 조합으로 구현되는 다양한 알고리즘을 포함하여, C, C++, 자바(Java), 어셈블러(assembler), 파이썬(Python) 등과 같은 프로그래밍 또는 스크립팅 언어로 구현될 수 있다. 기능적인 측면들은 하나 이상의 프로세서들에서 실행되는 알고리즘으로 구현될 수 있다.

[0107] 이상, 첨부된 도면을 참조로 하여 본 발명의 실시예를 설명하였지만, 본 발명이 속하는 기술분야의 통상의 기술자는 본 발명이 그 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 실시될 수 있다는 것을 이해할 수 있을 것이다. 그러므로, 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며, 제한적이 아닌 것으로 이해해야만 한다.

부호의 설명

[0108] 100 : 전자 장치

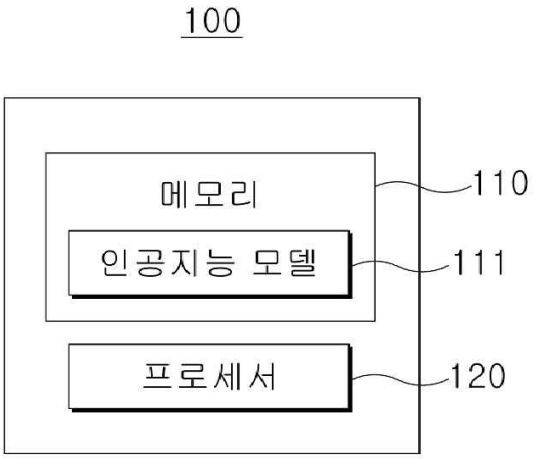
110 : 메모리

111 : 인공지능 모델

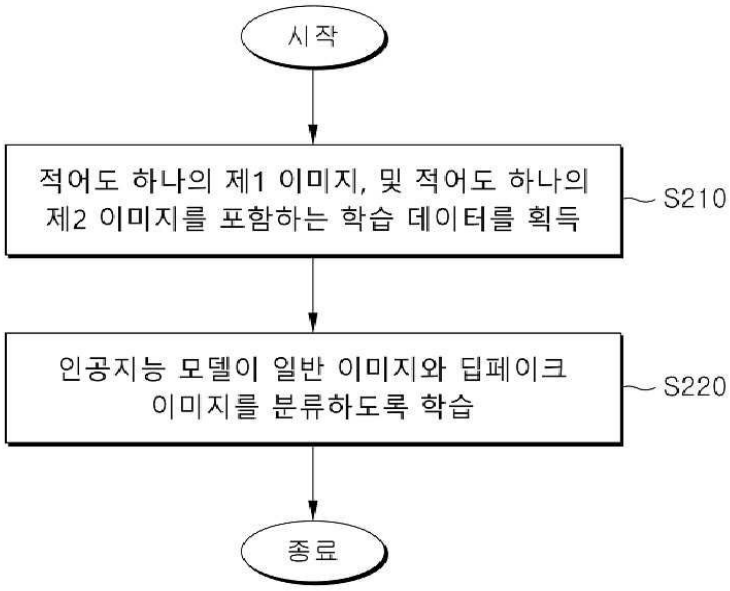
120 : 프로세서

도면

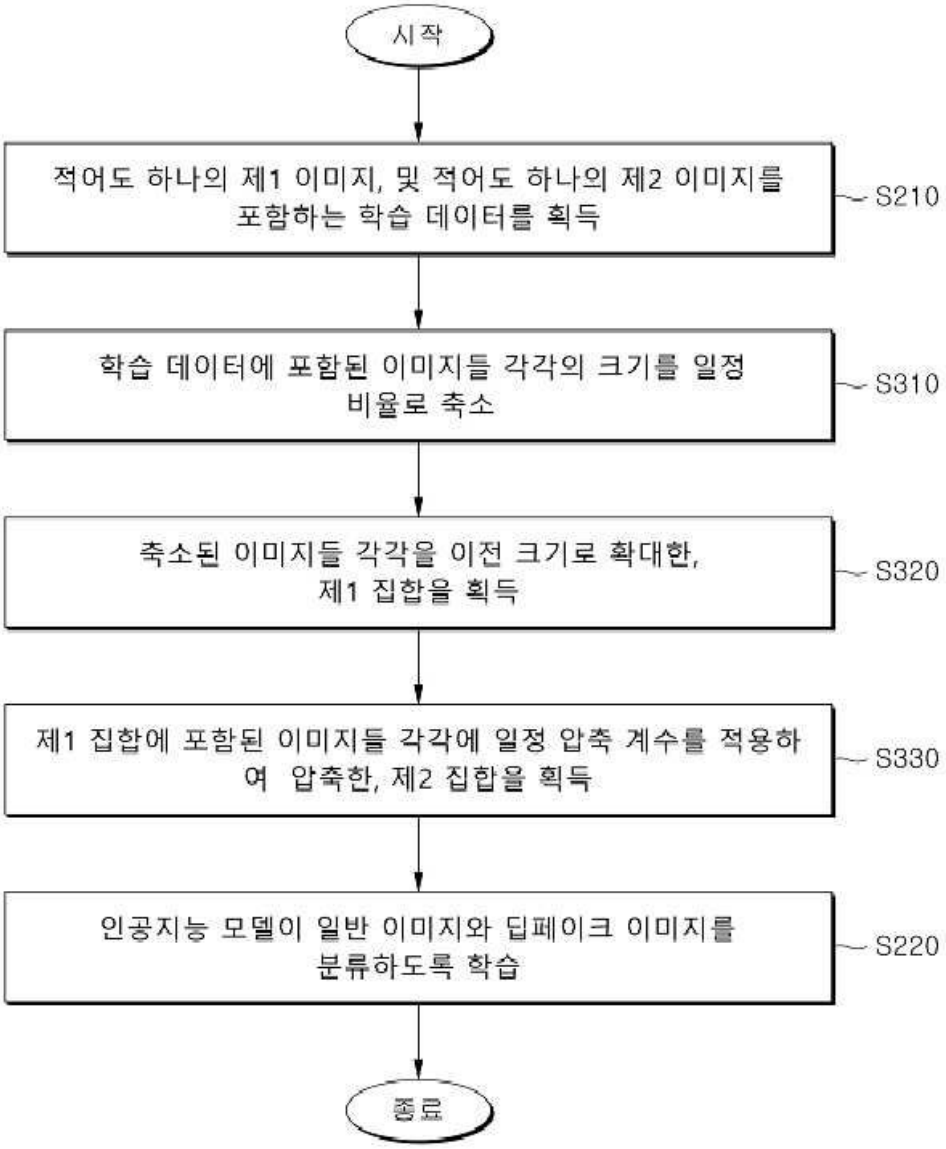
도면1



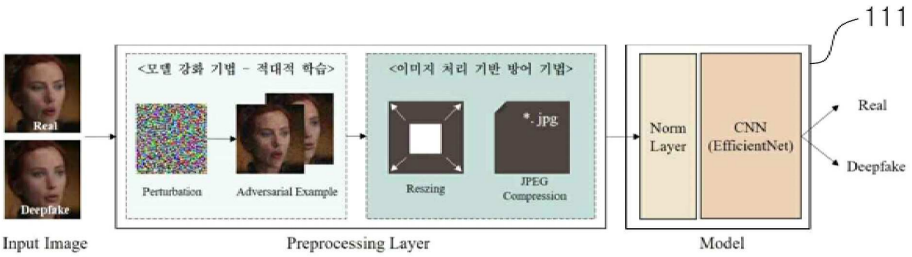
도면2


시작
적어도 하나의 제1 이미지, 및 적어도 하나의 제2 이미지를 포함하는 학습 데이터를 획득
S210
인공지능 모델이 일반 이미지와 딥페이크 이미지를 분류하도록 학습
S220
종료

도면3


시작
적어도 하나의 제1 이미지, 및 적어도 하나의 제2 이미지를 포함하는 학습 데이터를 획득
S210
학습 데이터에 포함된 이미지들 각각의 크기를 일정 비율로 축소
S310
축소된 이미지들 각각을 이전 크기로 확대한, 제1 집합을 획득
S320
제1 집합에 포함된 이미지들 각각에 일정 압축 계수를 적용하여 압축한, 제2 집합을 획득
S330
인공지능 모델이 일반 이미지와 딥페이크 이미지를 분류하도록 학습
S220
종료

도면4


111
Real
Deepfake
Input Image
<모델 강화 기법 - 적대적 학습>
Perturbation
Adversarial Example
<이미지 처리 기반 방어 기법>
Resizing
*.jpg
JPEG Compression
Preprocessing Layer
Norm Layer
CNN (EfficientNet)
Real
Deepfake
Model

도면5

ori:adv	orignal example		adversarial example	
	loss	acc	loss	acc
(baseline) 10:0	**0**	**100**	4.32	39.1
6:4	0.56	76.7	0.76	54.9
5:5	0.6	75.7	0.68	73.4
4:6	0.63	74.8	0.19	90.4
0:10	11.63	41	**0.07**	**96.1**

도면6

	orignal example		adversarial example	
	loss	acc	loss	acc
baseline (resize 0, jpeg 100)	0	100	4.32	39.1
adversarial training (4:6)	0.63	74.8	0.19	90.4
resize 3	1.05	76.1	0.24	91.3
resize 5	1.58	68	0.38	85.6
jpeg 5	0.68	82.6	0.36	85.2
jpeg 15	0.42	83.4	**0.17**	**93.8**
resize 3, jpeg 15	**0.35**	**85.4**	0.25	91.5

도면7

	orignal example		adversarial example	
	loss	acc	loss	acc
Baseline	0	100	4.32	39.1
Adversarial Training (4:6)	0.63	74.8	0.19	90.4
Resize 3, JPEG 15	0.35	85.4	0.25	91.5
FGSM & BIM & PGD	**0.22**	**96.4**	**0.02**	**98.4**