(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0095244
(43) 공개일자 2023년06월29일

(51) 국제특허분류(Int. Cl.)
G06T 3/40 (2006.01) *G06N 3/08* (2023.01)
G06T 7/00 (2017.01)
(52) CPC특허분류
G06T 3/40 (2013.01)
G06N 3/045 (2023.01)
(21) 출원번호 10-2021-0184526
(22) 출원일자 2021년12월22일
심사청구일자 2021년12월22일

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
허준범
경기도 용인시 기흥구 보정로 30, 119동 201호(보정동, 행원마을동아솔레시티아파트)
김민재
서울특별시 동대문구 왕산로 86, 오피스텔동 1312호(용두동, 동대문 푸르지오시티)
임경섭
서울특별시 동대문구 왕산로19바길 19(제기동)
(74) 대리인
이대호, 박건홍

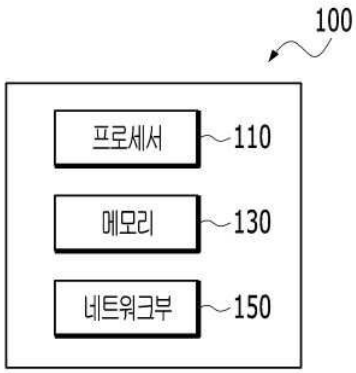
전체 청구항 수 : 총 12 항

(54) 발명의 명칭 **초해상화가 탑재된 물체 탐지 모델의 오작동을 유도하는 적대적 공격 방법**

(57) 요 약

본 개시의 몇몇 실시예에 따른 하나 이상의 프로세서를 포함하는 컴퓨팅 장치에 의해 적대적 공격을 수행하는 방법으로서, 원본 이미지를 제 1 신경망 모델에 입력하여 제 1 변환 이미지를 생성하는 단계; 상기 제 1 변환 이미지를 제 2 신경망 모델에 입력하여 제 1 물체 탐지 결과 데이터를 생성하는 단계; 상기 제 1 물체 탐지 결과 데이터와 기 저장된 실측 데이터(ground-truth) 간의 제 1 손실(loss) 값에 기초하여 제 1 노이즈(noise)를 생성하는 단계; 상기 제 1 노이즈 및 상기 제 1 변환 이미지에 기초하여 제 1 적대적 이미지를 생성하는 단계; 상기 제 1 적대적 이미지와 상기 제 1 변환 이미지 간의 제 2 손실 값에 기초하여 제 2 노이즈를 생성하는 단계; 및 상기 제 2 노이즈 및 상기 원본 이미지에 기초하여 제 2 적대적 이미지를 생성하는 단계;를 포함할 수 있다.

대 표 도 - 도1



(52) CPC특허분류
G06N 3/08 (2023.01)
G06T 7/00 (2013.01)
G06T 2207/20081 (2013.01)
G06T 2207/20084 (2013.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711126325
과제번호 2020-0-01819-002
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 정보통신방송혁신인재양성(R&D)
연구과제명 ICT명품인재양성(고려대학교)
기 여 율 25.0/100
과제수행기관명 고려대학교 산학협력단
연구기간 2021.01.01 ~ 2021.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1345342146
과제번호 2021R1A6A1A13044830
부처명 교육부
과제관리(전문)기관명 한국연구재단
연구사업명 이공학학술연구기반구축(R&D)
연구과제명 융합소프트웨어연구소
기 여 율 25.0/100
과제수행기관명 고려대학교 산학협력단
연구기간 2021.06.01 ~ 2022.02.28
이 발명을 지원한 국가연구개발사업
과제고유번호 1711139206
과제번호 2021-0-01810-001
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 정보통신방송혁신인재양성(R&D)
연구과제명 양자세계를 연결하는 초신뢰 양자인터넷 요소 기술 개발
기 여 율 25.0/100
과제수행기관명 고려대학교 산학협력단
연구기간 2021.07.01 ~ 2021.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1711152774
과제번호 2019-0-00533-004
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 정보보호핵심원천기술개발(정진기금)
연구과제명 컴퓨터 프로세서의 구조적 보안 취약점 검증 및 공격 탐지대응
기 여 율 25.0/100
과제수행기관명 고려대학교산학협력단
연구기간 2022.01.01 ~ 2022.12.31

명 세 서

청구범위

청구항 1

하나 이상의 프로세서를 포함하는 컴퓨팅 장치에 의해 적대적 공격을 수행하는 방법으로서,

원본 이미지를 제 1 신경망 모델에 입력하여 제 1 변환 이미지를 생성하는 단계;

상기 제 1 변환 이미지를 제 2 신경망 모델에 입력하여 제 1 물체 탐지 결과 데이터를 생성하는 단계;

상기 제 1 물체 탐지 결과 데이터와 기 저장된 실측 데이터(ground-truth) 간의 제 1 손실(loss) 값에 기초하여 제 1 노이즈(noise)를 생성하는 단계;

상기 제 1 노이즈 및 상기 제 1 변환 이미지에 기초하여 제 1 적대적 이미지를 생성하는 단계;

상기 제 1 적대적 이미지와 상기 제 1 변환 이미지 간의 제 2 손실 값에 기초하여 제 2 노이즈를 생성하는 단계; 및

상기 제 2 노이즈 및 상기 원본 이미지에 기초하여 제 2 적대적 이미지를 생성하는 단계;

를 포함하는,

적대적 공격을 수행하기 위한 방법.

청구항 2

제 1 항에 있어서,

제 1 신경망 모델은,

상기 원본 이미지에 기초하여 상기 원본 이미지보다 높은 해상도로 구성된 상기 제 1 변환 이미지를 생성하는 제 1 초해상화(super-resolution) 모델을 포함하는,

적대적 공격을 수행하기 위한 방법.

청구항 3

제 1 항에 있어서,

상기 제 2 신경망 모델은,

상기 제 1 변환 이미지에서 적어도 하나의 물체를 탐지하고, 그리고 적어도 하나의 물체의 위치 및 클래스(class)를 지정하여 상기 제 1 물체 탐지 결과 데이터를 생성하는 제 1 물체 탐지 모델을 포함하는,

적대적 공격을 수행하기 위한 방법.

청구항 4

제 1 항에 있어서,

상기 제 1 신경망 모델은 기 설정된 제 1 손실 함수(loss function)에 기초하여 사전 학습되고, 그리고

상기 제 1 물체 탐지 결과 데이터와 상기 기 저장된 실측 데이터 간의 상기 제 1 손실 값에 기초하여 상기 제 1 노이즈를 생성하는 단계는,

상기 기 설정된 제 1 손실 함수에 기초하여 상기 제 1 물체 탐지 결과 데이터와 상기 기 저장된 실측 데이터 간

의 상기 제 1 손실 값을 산출하는 단계; 및

산출된 상기 제 1 손실 값에 기초하여 상기 제 1 노이즈를 생성하는 단계;

를 포함하는,

적대적 공격을 수행하기 위한 방법.

청구항 5

제 1 항에 있어서,

상기 제 2 신경망 모델은 기 설정된 제 2 손실 함수에 기초하여 사전 학습되고, 그리고

상기 제 1 적대적 이미지와 상기 제 1 변환 이미지 간의 상기 제 2 손실 값에 기초하여 상기 제 2 노이즈를 생성하는 단계는,

상기 기 설정된 제 2 손실 함수에 기초하여 상기 제 1 적대적 이미지와 상기 제 1 변환 이미지 간의 상기 제 2 손실 값을 산출하는 단계; 및

산출된 상기 제 2 손실 값에 기초하여 상기 제 2 노이즈를 생성하는 단계;

를 포함하는,

적대적 공격을 수행하기 위한 방법.

청구항 6

제 1 항에 있어서,

상기 제 1 노이즈 및 상기 제 1 변환 이미지에 기초하여 상기 제 1 적대적 이미지를 생성하는 단계는,

상기 제 1 변환 이미지를 구성하는 적어도 하나의 제 1 변환 이미지 픽셀에 상기 제 1 노이즈를 추가하여 상기 제 1 적대적 이미지를 생성하는 단계;

를 포함하는,

적대적 공격을 수행하기 위한 방법.

청구항 7

제 1 항에 있어서,

상기 제 2 노이즈 및 상기 원본 이미지에 기초하여 상기 제 2 적대적 이미지를 생성하는 단계는,

상기 원본 이미지를 구성하는 적어도 하나의 원본 이미지 픽셀에 상기 제 2 노이즈를 추가하여 상기 제 2 적대적 이미지를 생성하는 단계;

를 포함하는,

적대적 공격을 수행하기 위한 방법.

청구항 8

제 1 항에 있어서,

상기 제 2 적대적 이미지를 제 3 신경망 모델에 입력하여 상기 제 3 신경망 모델의 성능을 판단하는 단계;

를 더 포함하는,

적대적 공격을 수행하기 위한 방법.

청구항 9

제 8 항에 있어서,

상기 제 2 적대적 이미지를 상기 제 3 신경망 모델에 입력하여 상기 제 3 신경망 모델의 성능을 판단하는 단계는,

상기 제 2 적대적 이미지를 상기 제 3 신경망 모델에 입력하여 제 2 물체 탐지 결과 데이터를 생성하는 단계; 및

상기 제 2 물체 탐지 결과 데이터와 상기 기 저장된 실측 데이터 간의 제 3 손실 값에 기초하여 상기 제 3 신경망 모델의 성능을 판단하는 단계;

를 포함하는,

적대적 공격을 수행하기 위한 방법.

청구항 10

제 9 항에 있어서,

상기 제 3 신경망 모델은,

상기 제 2 적대적 이미지에 기초하여 상기 제 2 적대적 이미지보다 높은 해상도로 구성된 제 2 변환 이미지를 생성하는 제 2 초해상화 모델; 및

상기 제 2 변환 이미지에서 적어도 하나의 물체를 탐지하고, 그리고 탐지된 적어도 하나의 물체의 위치 및 클래스를 지정하여 상기 제 2 물체 탐지 결과 데이터를 생성하는 제 2 물체 탐지 모델;

이 결합된 모델인,

적대적 공격을 수행하기 위한 방법.

청구항 11

컴퓨터 판독가능 저장 매체에 저장된 컴퓨터 프로그램으로서, 상기 컴퓨터 프로그램은 적대적 공격을 수행하기 위한 컴퓨팅 장치의 프로세서로 하여금 이하의 단계들을 수행하기 위한 명령들을 포함하며, 상기 단계들은:

원본 이미지를 제 1 신경망 모델에 입력하여 제 1 변환 이미지를 생성하는 단계;

상기 제 1 변환 이미지를 제 2 신경망 모델에 입력하여 제 1 물체 탐지 결과 데이터를 생성하는 단계;

상기 제 1 물체 탐지 결과 데이터와 기 저장된 실측 데이터(ground-truth) 간의 제 1 손실(loss) 값에 기초하여 제 1 노이즈(noise)를 생성하는 단계;

상기 제 1 노이즈 및 상기 제 1 변환 이미지에 기초하여 제 1 적대적 이미지를 생성하는 단계;

상기 제 1 적대적 이미지와 상기 제 1 변환 이미지 간의 제 2 손실 값에 기초하여 제 2 노이즈를 생성하는 단계; 및

상기 제 2 노이즈 및 상기 원본 이미지에 기초하여 제 2 적대적 이미지를 생성하는 단계;

를 포함하는,

컴퓨터 판독가능 저장 매체에 저장된 컴퓨터 프로그램.

청구항 12

적대적 공격을 수행하기 위한 컴퓨팅 장치에 있어서,

프로세서;

상기 프로세서에 의해 실행가능한 컴퓨터 프로그램을 저장하는 메모리; 및

네트워크부;

를 포함하고,

상기 프로세서는,

원본 이미지를 제 1 신경망 모델에 입력하여 제 1 변환 이미지를 생성하고,

상기 제 1 변환 이미지를 제 2 신경망 모델에 입력하여 제 1 물체 탐지 결과 데이터를 생성하고,

상기 제 1 물체 탐지 결과 데이터와 기 저장된 실측 데이터(ground-truth) 간의 제 1 손실(loss) 값에 기초하여 제 1 노이즈(noise)를 생성하고,

상기 제 1 노이즈 및 상기 제 1 변환 이미지에 기초하여 제 1 적대적 이미지를 생성하고,

상기 제 1 적대적 이미지와 상기 제 1 변환 이미지 간의 제 2 손실 값에 기초하여 제 2 노이즈를 생성하고, 그리고

상기 제 2 노이즈 및 상기 원본 이미지에 기초하여 제 2 적대적 이미지를 생성하는,

컴퓨팅 장치.

발명의 설명

기 술 분 야

[0001] 본 개시는 인공지능 기술에 관한 것으로, 보다 구체적으로, 초해상화가 탑재된 물체 탐지 모델의 오작동을 유도하는 적대적 공격 방법에 관한 것이다.

배 경 기 술

[0002] 물체 탐지 모델은 컴퓨터 비전(Computer vision)분야 중에서도 인공신경망 등을 이용하여 이미지에서 물체의 위치를 특정하고 해당 물체에 대한 분류(classification)를 진행하는 인공지능 모델이다.

[0003] 초해상화 모델은 인공신경망 등을 이용하여 원본 이미지의 화질과 최대한 유사하면서 원본보다 큰 해상도를 가지는 이미지를 생성하는 인공지능 모델이다.

[0004] 상술한 두 모델은 디지털 줌과 물체 탐지의 결합이나 위성사진에서의 작은 물체의 탐지에서 용이하게 사용될 수 있는 모델 조합으로, 컴퓨터 비전분야에서 가장 실용성이 높고 사용자들에게 노출되기 쉽다. 이는 곧 두 모델의 조합에 대한 적대적 공격(Adversarial attack)이 실제 사용자에게 미치는 영향이 치명적일 수 있다는 것을 의미하기도 한다.

[0005] 적대적 공격은 일반적으로 입력 데이터에 사람은 구분할 수 없는 노이즈(noise)를 추가하여 공격 대상이 되는 모델이 원래의 정상적인 결과를 도출하는 것을 방해하도록 하는 공격을 의미한다. 여기서 노이즈가 추가된 입력 데이터를 적대적 예제(Adversarial example)라 부를 수 있다. 그리고 컴퓨터 비전 모델에 대한 적대적 공격은 이미지의 각 픽셀에 노이즈를 일정 수준 추가하여 생성된 적대적 이미지(Adversarial image)를 통해 이루어진다.

선행기술문헌

특허문헌

[0006] (특허문헌 0001) 대한민국 등록특허 제10-2253402호(2021.05.12. 등록)

발명의 내용

해결하려는 과제

[0007] 본 개시는 전술한 배경기술에 대응하여 안출된 것으로, 초해상화가 탑재된 물체 탐지 모델의 오작동을 유도하는 적대적 공격 방법을 제공하고자 한다.

[0008] 본 개시의 기술적 과제들은 이상에서 언급한 기술적 과제로 제한되지 않으며, 언급되지 않은 또 다른 기술적 과제들은 아래의 기재로부터 당업자에게 명확하게 이해될 수 있을 것이다.

과제의 해결 수단

[0009] 전술한 바와 같은 과제를 해결하기 위한 본 개시의 몇몇 실시예에 따라, 하나 이상의 프로세서를 포함하는 컴퓨팅 장치에 의해 적대적 공격을 수행하는 방법으로서, 원본 이미지를 제 1 신경망 모델에 입력하여 제 1 변환 이미지를 생성하는 단계; 상기 제 1 변환 이미지를 제 2 신경망 모델에 입력하여 제 1 물체 탐지 결과 데이터를 생성하는 단계; 상기 제 1 물체 탐지 결과 데이터와 기 저장된 실측 데이터(ground-truth) 간의 제 1 손실(loss) 값에 기초하여 제 1 노이즈(noise)를 생성하는 단계; 상기 제 1 노이즈 및 상기 제 1 변환 이미지에 기초하여 제 1 적대적 이미지를 생성하는 단계; 상기 제 1 적대적 이미지와 상기 제 1 변환 이미지 간의 제 2 손실 값에 기초하여 제 2 노이즈를 생성하는 단계; 및 상기 제 2 노이즈 및 상기 원본 이미지에 기초하여 제 2 적대적 이미지를 생성하는 단계;를 포함할 수 있다.

[0010] 대안적으로, 제 1 신경망 모델은, 상기 원본 이미지에 기초하여 상기 원본 이미지보다 높은 해상도로 구성된 상기 제 1 변환 이미지를 생성하는 제 1 초해상화(super-resolution) 모델을 포함할 수 있다.

[0011] 대안적으로, 상기 제 2 신경망 모델은, 상기 제 1 변환 이미지에서 적어도 하나의 물체를 탐지하고, 그리고 적어도 하나의 물체의 위치 및 클래스(class)를 지정하여 상기 제 1 물체 탐지 결과 데이터를 생성하는 제 1 물체 탐지 모델을 포함할 수 있다.

[0012] 대안적으로, 상기 제 1 신경망 모델은 기 설정된 제 1 손실 함수(loss function)에 기초하여 사전 학습되고, 그리고 상기 제 1 물체 탐지 결과 데이터와 상기 기 저장된 실측 데이터 간의 상기 제 1 손실 값에 기초하여 상기 제 1 노이즈를 생성하는 단계는, 상기 기 설정된 제 1 손실 함수에 기초하여 상기 제 1 물체 탐지 결과 데이터와 상기 기 저장된 실측 데이터 간의 상기 제 1 손실 값을 산출하는 단계; 및 산출된 상기 제 1 손실 값에 기초하여 상기 제 1 노이즈를 생성하는 단계;를 포함할 수 있다.

[0013] 대안적으로, 상기 제 2 신경망 모델은 기 설정된 제 2 손실 함수에 기초하여 사전 학습되고, 그리고 상기 제 1 적대적 이미지와 상기 제 1 변환 이미지 간의 상기 제 2 손실 값에 기초하여 상기 제 2 노이즈를 생성하는 단계는, 상기 기 설정된 제 2 손실 함수에 기초하여 상기 제 1 적대적 이미지와 상기 제 1 변환 이미지 간의 상기 제 2 손실 값을 산출하는 단계; 및 산출된 상기 제 2 손실 값에 기초하여 상기 제 2 노이즈를 생성하는 단계;를 포함할 수 있다.

[0014] 대안적으로, 상기 제 1 노이즈 및 상기 제 1 변환 이미지에 기초하여 상기 제 1 적대적 이미지를 생성하는 단계는, 상기 제 1 변환 이미지를 구성하는 적어도 하나의 제 1 변환 이미지 픽셀에 상기 제 1 노이즈를 추가하여 상기 제 1 적대적 이미지를 생성하는 단계;를 포함할 수 있다.

[0015] 대안적으로, 상기 제 2 노이즈 및 상기 원본 이미지에 기초하여 상기 제 2 적대적 이미지를 생성하는 단계는, 상기 원본 이미지를 구성하는 적어도 하나의 원본 이미지 픽셀에 상기 제 2 노이즈를 추가하여 상기 제 2 적대적 이미지를 생성하는 단계;를 포함할 수 있다.

[0016] 대안적으로, 상기 제 2 적대적 이미지를 제 3 신경망 모델에 입력하여 상기 제 3 신경망 모델의 성능을 판단하는 단계;를 더 포함할 수 있다.

[0017] 대안적으로, 제 8 항에 있어서, 상기 제 2 적대적 이미지를 상기 제 3 신경망 모델에 입력하여 상기 제 3 신경망 모델의 성능을 판단하는 단계는, 상기 제 2 적대적 이미지를 상기 제 3 신경망 모델에 입력하여 제 2 물체 탐지 결과 데이터를 생성하는 단계; 및 상기 제 2 물체 탐지 결과 데이터와 상기 기 저장된 실측 데이터 간의

제 3 손실 값에 기초하여 상기 제 3 신경망 모델의 성능을 판단하는 단계;를 포함할 수 있다.

[0018] 대안적으로, 상기 제 3 신경망 모델은, 상기 제 2 적대적 이미지에 기초하여 상기 제 2 적대적 이미지보다 높은 해상도로 구성된 제 2 변환 이미지를 생성하는 제 2 초해상화 모델; 및 상기 제 2 변환 이미지에서 적어도 하나의 물체를 탐지하고, 그리고 탐지된 적어도 하나의 물체의 위치 및 클래스를 지정하여 상기 제 2 물체 탐지 결과 데이터를 생성하는 제 2 물체 탐지 모델;이 결합된 모델일 수 있다.

[0019] 전술한 바와 같은 과제를 해결하기 위한 본 개시의 몇몇 다른 실시예에 따라, 컴퓨터 판독가능 저장 매체에 저장된 컴퓨터 프로그램으로서, 상기 컴퓨터 프로그램은 적대적 공격을 수행하기 위한 컴퓨팅 장치의 프로세서로 하여금 이하의 단계들을 수행하기 위한 명령들을 포함하며, 상기 단계들은: 원본 이미지를 제 1 신경망 모델에 입력하여 제 1 변환 이미지를 생성하는 단계; 상기 제 1 변환 이미지를 제 2 신경망 모델에 입력하여 제 1 물체 탐지 결과 데이터를 생성하는 단계; 상기 제 1 물체 탐지 결과 데이터와 기 저장된 실측 데이터(ground-truth) 간의 제 1 손실(loss) 값에 기초하여 제 1 노이즈(noise)를 생성하는 단계; 상기 제 1 노이즈 및 상기 제 1 변환 이미지에 기초하여 제 1 적대적 이미지를 생성하는 단계; 상기 제 1 적대적 이미지와 상기 제 1 변환 이미지 간의 제 2 손실 값에 기초하여 제 2 노이즈를 생성하는 단계; 및 상기 제 2 노이즈 및 상기 원본 이미지에 기초하여 제 2 적대적 이미지를 생성하는 단계;를 포함할 수 있다.

[0020] 전술한 바와 같은 과제를 해결하기 위한 본 개시의 몇몇 또 다른 실시예에 따라, 적대적 공격을 수행하기 위한 컴퓨팅 장치에 있어서, 프로세서; 상기 프로세서에 의해 실행가능한 컴퓨터 프로그램을 저장하는 메모리; 및 네트워크부;를 포함하고, 상기 프로세서는, 원본 이미지를 제 1 신경망 모델에 입력하여 제 1 변환 이미지를 생성하고, 상기 제 1 변환 이미지를 제 2 신경망 모델에 입력하여 제 1 물체 탐지 결과 데이터를 생성하고, 상기 제 1 물체 탐지 결과 데이터와 기 저장된 실측 데이터(ground-truth) 간의 제 1 손실(loss) 값에 기초하여 제 1 노이즈(noise)를 생성하고, 상기 제 1 노이즈 및 상기 제 1 변환 이미지에 기초하여 제 1 적대적 이미지를 생성하고, 상기 제 1 적대적 이미지와 상기 제 1 변환 이미지 간의 제 2 손실 값에 기초하여 제 2 노이즈를 생성하고, 그리고 상기 제 2 노이즈 및 상기 원본 이미지에 기초하여 제 2 적대적 이미지를 생성할 수 있다.

발명의 효과

[0021] 본 개시는 초해상화가 탑재된 물체 탐지 모델의 오작동을 유도하는 적대적 공격을 수행할 수 있다.

[0022] 본 개시는 기존의 하나의 모델에 대해서만 공격을 수행하는 개별적 공격과 달리 2가지 모델에 대해서 공격을 수행하여 높은 성능을 가질 수 있다.

[0023] 본 개시는 초해상화가 탑재된 물체 탐지 모델에 적대적 공격을 수행하여 해당 모델의 평가를 통해 견고성(robustness)을 측정할 수 있다.

[0024] 본 개시는 컴퓨터 비전 모델의 방어 방법에 대한 연구의 시야를 넓힐 수 있다.

[0025] 본 개시에서 얻을 수 있는 효과는 이상에서 언급한 효과로 제한되지 않으며, 언급하지 않은 또 다른 효과들은 아래의 기재로부터 본 개시가 속하는 기술분야에서 통상의 지식을 가진 자에게 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

[0026] 다양한 양상들이 이제 도면들을 참조로 기재되며, 여기서 유사한 참조 번호들은 총괄적으로 유사한 구성요소들을 지칭하는데 이용된다. 이하의 실시예에서, 설명 목적을 위해, 다수의 특정 세부사항들이 하나 이상의 양상들의 총체적 이해를 제공하기 위해 제시된다. 그러나, 그러한 양상(들)이 이러한 구체적인 세부사항들 없이 실시될 수 있음은 명백할 것이다.

도 1은 본 개시의 몇몇 실시예에 따른 적대적 공격을 수행하는 방법을 제공하기 위한 컴퓨팅 장치의 블록 구성도이다.

도 2는 본 개시의 몇몇 실시예에 따른 신경망 모델을 나타낸 개략도이다.

도 3은 본 개시의 몇몇 실시예에 따른 적대적 공격을 수행하는 방법을 설명하기 위한 컴퓨팅 장치의 프로세서의 블록 구성도이다.

도 4는 본 개시의 몇몇 실시예에 따른 제 1 신경망 모델을 설명하기 위한 도면이다.

도 5는 본 개시의 몇몇 실시예에 따른 제 2 신경망 모델을 설명하기 위한 도면이다.

도 6은 본 개시의 몇몇 실시예에 따른 제 3 신경망 모델을 설명하기 위한 도면이다.

도 7은 본 개시의 몇몇 실시예에 따른 컴퓨팅 장치에 의해 수행되는 적대적 공격을 수행하는 방법을 설명하기 위한 순서도이다.

도 8은 본 개시의 실시예들이 구현될 수 있는 예시적인 컴퓨팅 환경에 대한 간략하고 일반적인 개략도를 도시한다.

발명을 실시하기 위한 구체적인 내용

[0027] 다양한 실시예들이 이제 도면을 참조하여 설명된다. 본 명세서에서, 다양한 설명들이 본 개시의 이해를 제공하기 위해서 제시된다. 그러나, 이러한 실시예들은 이러한 구체적인 설명 없이도 실행될 수 있음이 명백하다.

[0028] 본 명세서에서 사용되는 용어 "컴포넌트", "모듈", "시스템" 등은 컴퓨터-관련 엔티티, 하드웨어, 펌웨어, 소프트웨어, 소프트웨어 및 하드웨어의 조합, 또는 소프트웨어의 실행을 지칭한다. 예를 들어, 컴포넌트는 프로세서 상에서 실행되는 처리과정(procedure), 프로세서, 물체, 실행 스레드, 프로그램, 및/또는 컴퓨터일 수 있지만, 이들로 제한되는 것은 아니다. 예를 들어, 컴퓨팅 장치에서 실행되는 애플리케이션 및 컴퓨팅 장치 모두 컴포넌트일 수 있다. 하나 이상의 컴포넌트는 프로세서 및/또는 실행 스레드 내에 상주할 수 있다. 일 컴포넌트는 하나의 컴퓨터 내에 로컬화 될 수 있다. 일 컴포넌트는 2개 이상의 컴퓨터들 사이에 분배될 수 있다. 또한, 이러한 컴포넌트들은 그 내부에 저장된 다양한 데이터 구조들을 갖는 다양한 컴퓨터 판독가능한 매체로부터 실행할 수 있다. 컴포넌트들은 예를 들어 하나 이상의 데이터 패킷들을 갖는 신호(예를 들면, 로컬 시스템, 분산 시스템에서 다른 컴포넌트와 상호작용하는 하나의 컴포넌트로부터의 데이터 및/또는 신호를 통해 다른 시스템과 인터넷과 같은 네트워크를 통해 전송되는 데이터)에 따라 로컬 및/또는 원격 처리들을 통해 통신할 수 있다.

[0029] 더불어, 용어 "또는"은 배타적 "또는"이 아니라 내포적 "또는"을 의미하는 것으로 의도된다. 즉, 달리 특정되지 않거나 문맥상 명확하지 않은 경우에, "X는 A 또는 B를 이용한다"는 자연적인 내포적 치환 중 하나를 의미하는 것으로 의도된다. 즉, X가 A를 이용하거나; X가 B를 이용하거나; 또는 X가 A 및 B 모두를 이용하는 경우, "X는 A 또는 B를 이용한다"가 이들 경우들 어느 것으로도 적용될 수 있다. 또한, 본 명세서에 사용된 "및/또는"이라는 용어는 열거된 관련 아이템들 중 하나 이상의 아이템의 가능한 모든 조합을 지칭하고 포함하는 것으로 이해되어야 한다.

[0030] 또한, "포함한다" 및/또는 "포함하는"이라는 용어는, 해당 특징 및/또는 구성요소가 존재함을 의미하는 것으로 이해되어야 한다. 다만, "포함한다" 및/또는 "포함하는"이라는 용어는, 하나 이상의 다른 특징, 구성요소 및/또는 이들의 그룹의 존재 또는 추가를 배제하지 않는 것으로 이해되어야 한다. 또한, 달리 특정되지 않거나 단수 형태를 지시하는 것으로 문맥상 명확하지 않은 경우에, 본 명세서와 청구범위에서 단수는 일반적으로 "하나 또는 그 이상"을 의미하는 것으로 해석되어야 한다.

[0031] 그리고, "A 또는 B 중 적어도 하나"이라는 용어는, "A만을 포함하는 경우", "B 만을 포함하는 경우", "A와 B의 구성으로 조합된 경우"를 의미하는 것으로 해석되어야 한다.

[0032] 당업자들은 추가적으로 여기서 개시된 실시예들과 관련되어 설명된 다양한 예시적 논리적 블록들, 구성들, 모듈들, 회로들, 수단들, 로직들, 및 알고리즘 단계들이 전자 하드웨어, 컴퓨터 소프트웨어, 또는 양쪽 모두의 조합들로 구현될 수 있음을 인식해야 한다. 하드웨어 및 소프트웨어의 상호교환성을 명백하게 예시하기 위해, 다양한 예시적 컴포넌트들, 블록들, 구성들, 수단들, 로직들, 모듈들, 회로들, 및 단계들은 그들의 기능성 측면에서 일반적으로 위에서 설명되었다. 그러한 기능성이 하드웨어로 또는 소프트웨어로서 구현되는지 여부는 전반적인 시스템에 부과된 특정 어플리케이션(application) 및 설계 제한들에 달려 있다. 숙련된 기술자들은 각각의 특정 어플리케이션들을 위해 다양한 방법들로 설명된 기능성을 구현할 수 있다. 다만, 그러한 구현의 결정들이 본 개시내용의 영역을 벗어나게 하는 것으로 해석되어서는 안된다.

[0033] 제시된 실시예들에 대한 설명은 본 개시의 기술 분야에서 통상의 지식을 가진 자가 본 발명을 이용하거나 또는 실시할 수 있도록 제공된다. 이러한 실시예들에 대한 다양한 변형들은 본 개시의 기술 분야에서 통상의 지식을 가진 자에게 명백할 것이다. 여기에 정의된 일반적인 원리들은 본 개시의 범위를 벗어남이 없이 다른 실시예들에 적용될 수 있다. 그리하여, 본 발명은 여기에 제시된 실시예 들로 한정되는 것이 아니다. 본 발명은 여기에 제시된 원리들 및 신규한 특징들과 일관되는 최광의의 범위에서 해석되어야 할 것이다.

[0034] 본 개시에서 네트워크 함수와 인공 신경망 및 뉴럴 네트워크(neural network)는 상호교환이 가능하게 사용될 수 있다.

[0036] 도 1은 본 개시의 몇몇 실시예에 따른 적대적 공격을 수행하는 방법을 제공하기 위한 컴퓨팅 장치의 블록 구성도이다.

[0037] 도 1에 도시된 컴퓨팅 장치(100)의 구성은 간략화 하여 나타낸 예시일 뿐이다. 본 개시의 일 실시예에서 컴퓨팅 장치(100)는 컴퓨팅 장치(100)의 컴퓨팅 환경을 수행하기 위한 다른 구성들이 포함될 수 있고, 개시된 구성들 중 일부만이 컴퓨팅 장치(100)를 구성할 수도 있다.

[0038] 본 개시의 몇몇 실시예에 따른 컴퓨팅 장치(100)는 적대적 공격(Adversarial attack)을 수행하기 위한 장치일 수 있다. 적대적 공격은 특정 신경망 모델에 입력되는 입력 데이터에 노이즈(noise)를 추가하여 특정 신경망 모델이 원래의 정상적인 결과를 도출하는 것을 방해하는 공격일 수 있다.

[0039] 또한, 컴퓨팅 장치(100)는 적대적 공격을 수행하기 위해 신경망 모델들을 이용하여 적대적 이미지(Adversarial image)를 생성할 수 있다. 적대적 이미지는 특정 신경망 모델에 입력되는 이미지의 각 픽셀에 노이즈를 추가하여 생성된 이미지일 수 있다.

[0040] 그리고, 컴퓨팅 장치(100)는 적대적 이미지를 이용하여 평가 대상이 되는 신경망 모델의 성능을 판단할 수 있다.

[0041] 한편, 컴퓨팅 장치(100)는 프로세서(110), 메모리(130), 네트워크부(150)를 포함할 수 있다.

[0042] 프로세서(110)는 하나 이상의 코어로 구성될 수 있으며, 컴퓨팅 장치의 중앙 처리 장치(CPU: central processing unit), 범용 그래픽 처리 장치 (GPGPU: general purpose graphics processing unit), 텐서 처리 장치(TPU: tensor processing unit) 등의 데이터 분석, 딥러닝을 위한 프로세서를 포함할 수 있다. 프로세서(110)는 메모리(130)에 저장된 컴퓨터 프로그램을 판독하여 본 개시의 일 실시예에 따른 기계 학습을 위한 데이터 처리를 수행할 수 있다. 본 개시의 일실시예에 따라 프로세서(110)는 신경망의 학습을 위한 연산을 수행할 수 있다. 프로세서(110)는 딥러닝(DL: deep learning)에서 학습을 위한 입력 데이터의 처리, 입력 데이터에서의 피처 추출, 오차 계산, 역전파(backpropagation)를 이용한 신경망의 가중치 업데이트 등의 신경망의 학습을 위한 계산을 수행할 수 있다. 프로세서(110)의 CPU, GPGPU, 및 TPU 중 적어도 하나가 네트워크 함수의 학습을 처리할 수 있다. 예를 들어, CPU 와 GPGPU가 함께 네트워크 함수의 학습, 네트워크 함수를 이용한 데이터 분류를 처리할 수 있다. 또한, 본 개시의 일 실시예에서 복수의 컴퓨팅 장치의 프로세서를 함께 사용하여 네트워크 함수의 학습, 네트워크 함수를 이용한 데이터 분류를 처리할 수 있다. 또한, 본 개시의 일 실시예에 따른 컴퓨팅 장치에서 수행되는 컴퓨터 프로그램은 CPU, GPGPU 또는 TPU 실행가능 프로그램일 수 있다.

[0044] 본 개시의 몇몇 실시예에 따르면, 메모리(130)는 프로세서(110)가 생성하거나 결정한 임의의 형태의 정보 및 네트워크부(150)가 수신한 임의의 형태의 정보를 저장할 수 있다.

[0045] 본 개시의 몇몇 실시예에 따르면, 메모리(130)는 플래시 메모리 타입(flash memory type), 하드디스크 타입(hard disk type), 멀티미디어 카드 마이크로 타입(multimedia card micro type), 카드 타입의 메모리(예를 들어 SD 또는 XD 메모리 등), 램(Random Access Memory, RAM), SRAM(Static Random Access Memory), 롬(Read-Only Memory, ROM), EEPROM(Electrically Erasable Programmable Read-Only Memory), PROM(Programmable Read-Only Memory), 자기 메모리, 자기 디스크, 광디스크 중 적어도 하나의 타입의 저장매체를 포함할 수 있다. 컴퓨팅 장치(100)는 인터넷(internet) 상에서 상기 메모리(130)의 저장 기능을 수행하는 웹 스토리지(web storage)와 관련되어 동작할 수도 있다. 전술한 메모리에 대한 기재는 예시일 뿐, 본 개시는 이에 제한되지 않는다.

[0047] 본 개시의 몇몇 실시예에 따른 네트워크부(150)는 임의의 형태의 데이터 및 신호 등을 송수신할 수 있는 임의의 유무선 통신 네트워크가 본 개시 내용에서 표현되는 네트워크에 포함될 수 있다.

[0048] 본 명세서에서 설명된 기술들은 위에서 언급된 네트워크들뿐만 아니라, 다른 네트워크들에서도 사용될 수 있다.

[0050] 도 2는 본 개시의 몇몇 실시예에 따른 신경망 모델을 나타낸 개략도이다.

[0051] 본 명세서에 걸쳐, 연산 모델, 신경망, 신경망 모델, 네트워크 함수, 뉴럴 네트워크(neural network)는 동일한 의미로 사용될 수 있다. 신경망은 일반적으로 노드라 지칭될 수 있는 상호 연결된 계산 단위들의 집합으로 구성될 수 있다. 이러한 노드들은 뉴런(neuron)들로 지칭될 수도 있다. 신경망은 적어도 하나 이상의 노드들을 포함하여 구성된다. 신경망들을 구성하는 노드(또는 뉴런)들은 하나 이상의 링크에 의해 상호 연결될 수 있다.

[0052] 신경망 내에서, 링크를 통해 연결된 하나 이상의 노드들은 상대적으로 입력 노드 및 출력 노드의 관계를 형성할 수 있다. 입력 노드 및 출력 노드의 개념은 상대적인 것으로서, 하나의 노드에 대하여 출력 노드 관계에 있는 임의의 노드는 다른 노드와의 관계에서 입력 노드 관계에 있을 수 있으며, 그 역도 성립할 수 있다. 상술한 바와 같이, 입력 노드 대 출력 노드 관계는 링크를 중심으로 생성될 수 있다. 하나의 입력 노드에 하나 이상의 출력 노드가 링크를 통해 연결될 수 있으며, 그 역도 성립할 수 있다.

[0053] 하나의 링크를 통해 연결된 입력 노드 및 출력 노드 관계에서, 출력 노드의 데이터는 입력 노드에 입력된 데이터에 기초하여 그 값이 결정될 수 있다. 여기서 입력 노드와 출력 노드를 상호 연결하는 링크는 가중치(weight)를 가질 수 있다. 가중치는 가변적일 수 있으며, 신경망이 원하는 기능을 수행하기 위해, 사용자 또는 알고리즘에 의해 가변 될 수 있다. 예를 들어, 하나의 출력 노드에 하나 이상의 입력 노드가 각각의 링크에 의해 상호 연결된 경우, 출력 노드는 상기 출력 노드와 연결된 입력 노드들에 입력된 값들 및 각각의 입력 노드들에 대응하는 링크에 설정된 가중치에 기초하여 출력 노드 값을 결정할 수 있다.

[0054] 상술한 바와 같이, 신경망은 하나 이상의 노드들이 하나 이상의 링크를 통해 상호 연결되어 신경망 내에서 입력 노드 및 출력 노드 관계를 형성한다. 신경망 내에서 노드들과 링크들의 개수 및 노드들과 링크들 사이의 연관관계, 링크들 각각에 부여된 가중치의 값에 따라, 신경망의 특성이 결정될 수 있다. 예를 들어, 동일한 개수의 노드 및 링크들이 존재하고, 링크들의 가중치 값이 상이한 두 신경망이 존재하는 경우, 두 개의 신경망들은 서로 상이한 것으로 인식될 수 있다.

[0055] 신경망은 하나 이상의 노드들의 집합으로 구성될 수 있다. 신경망을 구성하는 노드들의 부분 집합은 레이어(layer)를 구성할 수 있다. 신경망을 구성하는 노드들 중 일부는, 최초 입력 노드로부터의 거리들에 기초하여, 하나의 레이어(layer)를 구성할 수 있다. 예를 들어, 최초 입력 노드로부터 거리가 n인 노드들의 집합은, n 레이어를 구성할 수 있다. 최초 입력 노드로부터 거리는, 최초 입력 노드로부터 해당 노드까지 도달하기 위해 거쳐야 하는 링크들의 최소 개수에 의해 정의될 수 있다. 그러나, 이러한 레이어의 정의는 설명을 위한 임의적인 것으로서, 신경망 내에서 레이어의 차수는 상술한 것과 상이한 방법으로 정의될 수 있다. 예를 들어, 노드들의 레이어는 최종 출력 노드로부터 거리에 의해 정의될 수도 있다.

[0056] 최초 입력 노드는 신경망 내의 노드들 중 다른 노드들과의 관계에서 링크를 거치지 않고 데이터가 직접 입력되는 하나 이상의 노드들을 의미할 수 있다. 또는, 신경망 네트워크 내에서, 링크를 기준으로 한 노드 간의 관계에 있어서, 링크로 연결된 다른 입력 노드들을 가지지 않는 노드들을 의미할 수 있다. 이와 유사하게, 최종 출력 노드는 신경망 내의 노드들 중 다른 노드들과의 관계에서, 출력 노드를 가지지 않는 하나 이상의 노드들을 의미할 수 있다. 또한, 히든 노드는 최초 입력 노드 및 최후 출력 노드가 아닌 신경망을 구성하는 노드들을 의미할 수 있다.

[0057] 본 개시의 일 실시예에 따른 신경망은 입력 레이어의 노드의 개수가 출력 레이어의 노드의 개수와 동일할 수 있으며, 입력 레이어에서 히든 레이어로 진행됨에 따라 노드의 수가 감소하다가 다시 증가하는 형태의 신경망일 수 있다. 또한, 본 개시의 다른 일 실시예에 따른 신경망은 입력 레이어의 노드의 개수가 출력 레이어의 노드의 개수 보다 적을 수 있으며, 입력 레이어에서 히든 레이어로 진행됨에 따라 노드의 수가 감소하는 형태의 신경망일 수 있다. 또한, 본 개시의 또 다른 일 실시예에 따른 신경망은 입력 레이어의 노드의 개수가 출력 레이어의 노드의 개수보다 많을 수 있으며, 입력 레이어에서 히든 레이어로 진행됨에 따라 노드의 수가 증가하는 형태의 신경망일 수 있다. 본 개시의 또 다른 일 실시예에 따른 신경망은 상술한 신경망들의 조합된 형태의 신경망일 수 있다.

[0058] 딥 뉴럴 네트워크(DNN: deep neural network, 심층신경망)는 입력 레이어와 출력 레이어 외에 복수의 히든 레이어를 포함하는 신경망을 의미할 수 있다. 딥 뉴럴 네트워크를 이용하면 데이터의 잠재적인 구조(latent

structures)를 파악할 수 있다. 즉, 사진, 글, 비디오, 음성, 음악의 잠재적인 구조(예를 들어, 어떤 물체가 사진에 있는지, 글의 내용과 감정이 무엇인지, 음성의 내용과 감정이 무엇인지 등)를 파악할 수 있다. 딥 뉴럴 네트워크는 컨볼루션 뉴럴 네트워크(CNN: convolutional neural network), 리커런트 뉴럴 네트워크(RNN: recurrent neural network), 오토 인코더(auto encoder), 제한 볼츠만 머신(RBM: restricted boltzmann machine), 심층 신뢰 네트워크(DBN: deep belief network), Q 네트워크, U 네트워크, 샴 네트워크, 적대적 생성 네트워크(GAN: Generative Adversarial Network) 등을 포함할 수 있다. 전술한 딥 뉴럴 네트워크의 기재는 예시일 뿐이며 본 개시는 이에 제한되지 않는다.

[0059] 본 개시의 일 실시예에서 네트워크 함수는 오토 인코더(autoencoder)를 포함할 수도 있다. 오토 인코더는 입력 데이터와 유사한 출력 데이터를 출력하기 위한 인공 신경망의 일종일 수 있다. 오토 인코더는 적어도 하나의 히든 레이어를 포함할 수 있으며, 홀수 개의 히든 레이어가 입출력 레이어 사이에 배치될 수 있다. 각각의 레이어의 노드의 수는 입력 레이어의 노드의 수에서 병목 레이어(인코딩)라는 중간 레이어로 축소되었다가, 병목 레이어에서 출력 레이어(입력 레이어와 대칭)로 축소와 대칭되어 확장될 수도 있다. 오토 인코더는 비선형 차원 감소를 수행할 수 있다. 입력 레이어 및 출력 레이어의 수는 입력 데이터의 전처리 이후에 차원과 대응될 수 있다. 오토 인코더 구조에서 인코더에 포함된 히든 레이어의 노드의 수는 입력 레이어에서 멀어질수록 감소하는 구조를 가질 수 있다. 병목 레이어(인코더와 디코더 사이에 위치하는 가장 적은 노드를 가진 레이어)의 노드의 수는 너무 작은 경우 충분한 양의 정보가 전달되지 않을 수 있으므로, 특정 수 이상(예를 들어, 입력 레이어의 절반 이상 등)으로 유지될 수도 있다.

[0060] 뉴럴 네트워크는 교사 학습(supervised learning), 비교사 학습(unsupervised learning), 반교사학습(semi supervised learning), 또는 강화학습(reinforcement learning) 중 적어도 하나의 방식으로 학습될 수 있다. 뉴럴 네트워크의 학습은 뉴럴 네트워크가 특정한 동작을 수행하기 위한 지식을 뉴럴 네트워크에 적용하는 과정일 수 있다.

[0061] 뉴럴 네트워크는 출력의 오류를 최소화하는 방향으로 학습될 수 있다. 뉴럴 네트워크의 학습에서 반복적으로 학습 데이터를 뉴럴 네트워크에 입력시키고 학습 데이터에 대한 뉴럴 네트워크의 출력과 타겟의 에러를 계산하고, 에러를 줄이기 위한 방향으로 뉴럴 네트워크의 에러를 뉴럴 네트워크의 출력 레이어에서부터 입력 레이어 방향으로 역전파(backpropagation)하여 뉴럴 네트워크의 각 노드의 가중치를 업데이트 하는 과정이다. 교사 학습의 경우 각각의 학습 데이터에 정답이 라벨링되어있는 학습 데이터를 사용하며(즉, 라벨링된 학습 데이터), 비교사 학습의 경우는 각각의 학습 데이터에 정답이 라벨링되어 있지 않을 수 있다. 즉, 예를 들어 데이터 분류에 관한 교사 학습의 경우의 학습 데이터는 학습 데이터 각각에 카테고리가 라벨링 된 데이터 일 수 있다. 라벨링된 학습 데이터가 뉴럴 네트워크에 입력되고, 뉴럴 네트워크의 출력(카테고리)과 학습 데이터의 라벨을 비교함으로써 오류(error)가 계산될 수 있다. 다른 예로, 데이터 분류에 관한 비교사 학습의 경우 입력인 학습 데이터가 뉴럴 네트워크 출력과 비교됨으로써 오류가 계산될 수 있다. 계산된 오류는 뉴럴 네트워크에서 역방향(즉, 출력 레이어에서 입력 레이어 방향)으로 역전파 되며, 역전파에 따라 뉴럴 네트워크의 각 레이어의 각 노드들의 연결 가중치가 업데이트 될 수 있다. 업데이트 되는 각 노드의 연결 가중치는 학습률(learning rate)에 따라 변화량이 결정될 수 있다. 입력 데이터에 대한 뉴럴 네트워크의 계산과 에러의 역전파는 학습 사이클(epoch)을 구성할 수 있다. 학습률은 뉴럴 네트워크의 학습 사이클의 반복 횟수에 따라 상이하게 적용될 수 있다. 예를 들어, 뉴럴 네트워크의 학습 초기에는 높은 학습률을 사용하여 뉴럴 네트워크가 빠르게 일정 수준의 성능을 확보하도록 하여 효율성을 높이고, 학습 후기에는 낮은 학습률을 사용하여 정확도를 높일 수 있다.

[0062] 뉴럴 네트워크의 학습에서 일반적으로 학습 데이터는 실제 데이터(즉, 학습된 뉴럴 네트워크를 이용하여 처리하고자 하는 데이터)의 부분집합일 수 있으며, 따라서, 학습 데이터에 대한 오류는 감소하나 실제 데이터에 대해서는 오류가 증가하는 학습 사이클이 존재할 수 있다. 과적합(overfitting)은 이와 같이 학습 데이터에 과하게 학습하여 실제 데이터에 대한 오류가 증가하는 현상이다. 예를 들어, 노란색 고양이를 보여 고양이를 학습한 뉴럴 네트워크가 노란색 이외의 고양이를 보고는 고양이임을 인식하지 못하는 현상이 과적합의 일종일 수 있다. 과적합은 머신러닝 알고리즘의 오류를 증가시키는 원인으로 작용할 수 있다. 이러한 과적합을 막기 위하여 다양한 최적화 방법이 사용될 수 있다. 과적합을 막기 위해서는 학습 데이터를 증가시키거나, 레귤라이제이션(regularization), 학습의 과정에서 네트워크의 노드 일부를 비활성화하는 드롭아웃(dropout), 배치 정규화 레이어(batch normalization layer)의 활용 등의 방법이 적용될 수 있다.

[0064] 본 개시의 일 실시예에 따라 데이터 구조를 저장한 컴퓨터 판독가능 매체가 개시된다.

[0065] 데이터 구조는 데이터에 효율적인 접근 및 수정을 가능하게 하는 데이터의 조직, 관리, 저장을 의미할 수 있다. 데이터 구조는 특정 문제(예를 들어, 최단 시간으로 데이터 검색, 데이터 저장, 데이터 수정) 해결을 위한 데이터의 조직을 의미할 수 있다. 데이터 구조는 특정한 데이터 처리 기능을 지원하도록 설계된, 데이터 요소들 간의 물리적이거나 논리적인 관계로 정의될 수도 있다. 데이터 요소들 간의 논리적인 관계는 사용자 정의 데이터 요소들 간의 연결관계를 포함할 수 있다. 데이터 요소들 간의 물리적인 관계는 컴퓨터 판독가능 저장매체(예를 들어, 영구 저장 장치)에 물리적으로 저장되어 있는 데이터 요소들 간의 실제 관계를 포함할 수 있다. 데이터 구조는 구체적으로 데이터의 집합, 데이터 간의 관계, 데이터에 적용할 수 있는 함수 또는 명령어를 포함할 수 있다. 효과적으로 설계된 데이터 구조를 통해 컴퓨팅 장치는 컴퓨팅 장치의 자원을 최소한으로 사용하면서 연산을 수행할 수 있다. 구체적으로 컴퓨팅 장치는 효과적으로 설계된 데이터 구조를 통해 연산, 읽기, 삽입, 삭제, 비교, 교환, 검색의 효율성을 높일 수 있다.

[0066] 데이터 구조는 데이터 구조의 형태에 따라 선형 데이터 구조와 비선형 데이터 구조로 구분될 수 있다. 선형 데이터 구조는 하나의 데이터 뒤에 하나의 데이터만이 연결되는 구조일 수 있다. 선형 데이터 구조는 리스트(List), 스택(Stack), 큐(Queue), 데크(Deque)를 포함할 수 있다. 리스트는 내부적으로 순서가 존재하는 일련의 데이터 집합을 의미할 수 있다. 리스트는 연결 리스트(Linked List)를 포함할 수 있다. 연결 리스트는 각각의 데이터가 포인터를 가지고 한 줄로 연결되어 있는 방식으로 데이터가 연결된 데이터 구조일 수 있다. 연결 리스트에서 포인터는 다음이나 이전 데이터와의 연결 정보를 포함할 수 있다. 연결 리스트는 형태에 따라 단일 연결 리스트, 이중 연결 리스트, 원형 연결 리스트로 표현될 수 있다. 스택은 제한적으로 데이터에 접근할 수 있는 데이터 나열 구조일 수 있다. 스택은 데이터 구조의 한 쪽 끝에서만 데이터를 처리(예를 들어, 삽입 또는 삭제)할 수 있는 선형 데이터 구조일 수 있다. 스택에 저장된 데이터는 늦게 들어갈수록 빨리 나오는 데이터 구조(LIFO-Last in First Out)일 수 있다. 큐는 제한적으로 데이터에 접근할 수 있는 데이터 나열 구조로서, 스택과 달리 늦게 저장된 데이터일수록 늦게 나오는 데이터 구조(FIFO-First in First Out)일 수 있다. 데크는 데이터 구조의 양 쪽 끝에서 데이터를 처리할 수 있는 데이터 구조일 수 있다.

[0067] 비선형 데이터 구조는 하나의 데이터 뒤에 복수개의 데이터가 연결되는 구조일 수 있다. 비선형 데이터 구조는 그래프(Graph) 데이터 구조를 포함할 수 있다. 그래프 데이터 구조는 정점(Vertex)과 간선(Edge)으로 정의될 수 있으며 간선은 서로 다른 두개의 정점을 연결하는 선을 포함할 수 있다. 그래프 데이터 구조 트리(Tree) 데이터 구조를 포함할 수 있다. 트리 데이터 구조는 트리에 포함된 복수개의 정점 중에서 서로 다른 두개의 정점을 연결시키는 경로가 하나인 데이터 구조일 수 있다. 즉 그래프 데이터 구조에서 루프(loop)를 형성하지 않는 데이터 구조일 수 있다.

[0068] 본 명세서에 걸쳐, 연산 모델, 신경망, 신경망 모델, 네트워크 함수, 뉴럴 네트워크(neural network)는 동일한 의미로 사용될 수 있다. 이하에서는 신경망으로 통일하여 기술한다. 데이터 구조는 신경망을 포함할 수 있다. 그리고 신경망을 포함한 데이터 구조는 컴퓨터 판독가능 매체에 저장될 수 있다. 신경망을 포함한 데이터 구조는 또한 신경망에 의한 처리를 위하여 전처리된 데이터, 신경망에 입력되는 데이터, 신경망의 가중치, 신경망의 하이퍼 파라미터, 신경망으로부터 획득한 데이터, 신경망의 각 노드 또는 레이어와 연관된 활성 함수, 신경망의 학습을 위한 손실 함수 등을 포함할 수 있다. 신경망을 포함한 데이터 구조는 상기 개시된 구성들 중 임의의 구성 요소들을 포함할 수 있다. 즉 신경망을 포함한 데이터 구조는 신경망에 의한 처리를 위하여 전처리된 데이터, 신경망에 입력되는 데이터, 신경망의 가중치, 신경망의 하이퍼 파라미터, 신경망으로부터 획득한 데이터, 신경망의 각 노드 또는 레이어와 연관된 활성 함수, 신경망의 학습을 위한 손실 함수 등 전부 또는 이들의 임의의 조합을 포함하여 구성될 수 있다. 전술한 구성들 이외에도, 신경망을 포함한 데이터 구조는 신경망의 특성을 결정하는 임의의 다른 정보를 포함할 수 있다. 또한, 데이터 구조는 신경망의 연산 과정에 사용되거나 발생되는 모든 형태의 데이터를 포함할 수 있으며 전술한 사항에 제한되는 것은 아니다. 컴퓨터 판독가능 매체는 컴퓨터 판독가능 기록 매체 및/또는 컴퓨터 판독가능 전송 매체를 포함할 수 있다. 신경망은 일반적으로 노드라 지칭될 수 있는 상호 연결된 계산 단위들의 집합으로 구성될 수 있다. 이러한 노드들은 뉴런(neuron)들로 지칭될 수도 있다. 신경망은 적어도 하나 이상의 노드들을 포함하여 구성된다.

[0069] 데이터 구조는 신경망에 입력되는 데이터를 포함할 수 있다. 신경망에 입력되는 데이터를 포함하는 데이터 구조는 컴퓨터 판독가능 매체에 저장될 수 있다. 신경망에 입력되는 데이터는 신경망 학습 과정에서 입력되는 학습 데이터 및/또는 학습이 완료된 신경망에 입력되는 입력 데이터를 포함할 수 있다. 신경망에 입력되는 데이터는 전처리(pre-processing)를 거친 데이터 및/또는 전처리 대상이 되는 데이터를 포함할 수 있다. 전처리는 데이터를 신경망에 입력시키기 위한 데이터 처리 과정을 포함할 수 있다. 따라서 데이터 구조는 전처리 대상이 되는 데이터 및 전처리로 발생되는 데이터를 포함할 수 있다. 전술한 데이터 구조는 예시일 뿐 본 개시는 이에 제한

되지 않는다.

[0070] 데이터 구조는 신경망의 가중치를 포함할 수 있다. (본 명세서에서 가중치, 파라미터는 동일한 의미로 사용될 수 있다.) 그리고 신경망의 가중치를 포함한 데이터 구조는 컴퓨터 판독가능 매체에 저장될 수 있다. 신경망은 복수개의 가중치를 포함할 수 있다. 가중치는 가변적일 수 있으며, 신경망이 원하는 기능을 수행하기 위해, 사용자 또는 알고리즘에 의해 가변 될 수 있다. 예를 들어, 하나의 출력 노드에 하나 이상의 입력 노드가 각각의 링크에 의해 상호 연결된 경우, 출력 노드는 상기 출력 노드와 연결된 입력 노드들에 입력된 값들 및 각각의 입력 노드들에 대응하는 링크에 설정된 가중치에 기초하여 출력 노드에서 출력되는 데이터 값을 결정할 수 있다. 전술한 데이터 구조는 예시일 뿐 본 개시는 이에 제한되지 않는다.

[0071] 제한이 아닌 예로서, 가중치는 신경망 학습 과정에서 가변되는 가중치 및/또는 신경망 학습이 완료된 가중치를 포함할 수 있다. 신경망 학습 과정에서 가변되는 가중치는 학습 사이클이 시작되는 시점의 가중치 및/또는 학습 사이클 동안 가변되는 가중치를 포함할 수 있다. 신경망 학습이 완료된 가중치는 학습 사이클이 완료된 가중치를 포함할 수 있다. 따라서 신경망의 가중치를 포함한 데이터 구조는 신경망 학습 과정에서 가변되는 가중치 및/또는 신경망 학습이 완료된 가중치를 포함한 데이터 구조를 포함할 수 있다. 그러므로 상술한 가중치 및/또는 각 가중치의 조합은 신경망의 가중치를 포함한 데이터 구조에 포함되는 것으로 한다. 전술한 데이터 구조는 예시일 뿐 본 개시는 이에 제한되지 않는다.

[0072] 신경망의 가중치를 포함한 데이터 구조는 직렬화(serialization) 과정을 거친 후 컴퓨터 판독가능 저장 매체(예를 들어, 메모리, 하드 디스크)에 저장될 수 있다. 직렬화는 데이터 구조를 동일하거나 다른 컴퓨팅 장치에 저장하고 나중에 다시 재구성하여 사용할 수 있는 형태로 변환하는 과정일 수 있다. 컴퓨팅 장치는 데이터 구조를 직렬화하여 네트워크를 통해 데이터를 송수신할 수 있다. 직렬화된 신경망의 가중치를 포함한 데이터 구조는 역직렬화(deserialization)를 통해 동일한 컴퓨팅 장치 또는 다른 컴퓨팅 장치에서 재구성될 수 있다. 신경망의 가중치를 포함한 데이터 구조는 직렬화에 한정되는 것은 아니다. 나아가 신경망의 가중치를 포함한 데이터 구조는 컴퓨팅 장치의 자원을 최소한으로 사용하면서 연산의 효율을 높이기 위한 데이터 구조(예를 들어, 비선형 데이터 구조에서 B-Tree, Trie, m-way search tree, AVL tree, Red-Black Tree)를 포함할 수 있다. 전술한 사항은 예시일 뿐 본 개시는 이에 제한되지 않는다.

[0073] 데이터 구조는 신경망의 하이퍼 파라미터(Hyper-parameter)를 포함할 수 있다. 그리고 신경망의 하이퍼 파라미터를 포함한 데이터 구조는 컴퓨터 판독가능 매체에 저장될 수 있다. 하이퍼 파라미터는 사용자에 의해 가변되는 변수일 수 있다. 하이퍼 파라미터는 예를 들어, 학습률(learning rate), 비용 함수(cost function), 학습 사이클 반복 횟수, 가중치 초기화(Weight initialization)(예를 들어, 가중치 초기화 대상이 되는 가중치 값의 범위 설정), Hidden Unit 개수(예를 들어, 히든 레이어의 개수, 히든 레이어의 노드 수)를 포함할 수 있다. 전술한 데이터 구조는 예시일 뿐 본 개시는 이에 제한되지 않는다.

[0075] 앞서 도 1 및 2를 통해 서술한 신경망 모델을 이용하여 적대적 공격을 수행하기 위한 프로세서(110)의 구성에 대해서 도 3을 참조하여 후술한다.

[0076] 도 3은 본 개시의 몇몇 실시예에 따른 적대적 공격을 수행하는 방법을 설명하기 위한 컴퓨팅 장치의 프로세서의 블록 구성도이다.

[0077] 도 3을 참조하면, 컴퓨팅 장치(100)의 프로세서(110)는 제 1 신경망 모델(200), 제 2 신경망 모델(300), 제 1 노이즈 생성부(400), 제 1 적대적 이미지 생성부(500), 제 2 노이즈 생성부(600), 제 2 적대적 이미지 생성부(700) 및 제 3 신경망 모델(800)을 포함할 수 있다. 다만, 상술한 구성 요소들을 프로세서(110)를 구현하는데 있어서 필수적인 것은 아니며, 프로세서(110)는 위에 열거된 구성 요소들보다 많거나, 또는 적은 구성요소들을 가질 수 있다.

[0078] 제 1 신경망 모델(200)은 원본 이미지가 입력될 수 있고, 입력된 원본 이미지를 변환하여 제 1 변환 이미지를 생성할 수 있다. 제 1 신경망 모델(200)은 원본 이미지에 기초하여 원본 이미지보다 높은 해상도로 구성된 제 1 변환 이미지를 생성하는 제 1 초해상화(super-resolution) 모델을 포함할 수 있다.

[0079] 제 1 신경망 모델(200)은 SRCNN(Super-Resolution Convolutional Neural Network), RCAN(Residual channel attention network) 및/또는 DBPN(Deep Back-Projection Network) 중 적어도 하나를 포함하는 제 1 초해상화 모델을 포함할 수 있다.

[0080] 여기서, SRCNN은 복수의 층(예를 들어, 세 개의 층)으로 구성된 신경망 모델일 수 있다. SRCNN은 원본 이미지를 전처리 단계에서 바이큐빅(bicubic) 방식으로 다운 샘플링하여 저해상도 이미지를 생성하는 신경망 모델일 수 있다. 그리고, SRCNN은 저해상도 이미지에서부터 생성된 선명하지 않은 이미지와 원본 이미지 사이의 특징을 학습하는 방식으로 사전 학습될 수 있다. SRCNN은 입력된 원본 이미지를 원본 이미지보다 높은 해상도로 구성된 제 1 변환 이미지를 생성하는 신경망 모델일 수 있다.

[0081] RCAN은 컨볼루션 레이어를 통해 원본 이미지의 특징 정보를 추출하는 신경망 모델일 수 있다. 그리고, RCAN은 channel attention module을 통해 특징 정보에서 심층 특징 정보를 추출하여 고해상도 이미지인 제 1 변환 이미지를 생성하는 신경망 모델일 수 있다.

[0082] DBPN은 도 4를 참조하여 후술하도록 한다. 도 4는 본 개시의 몇몇 실시예에 따른 제 1 신경망 모델을 설명하기 위한 도면이다.

[0083] 도 4를 참조하면, DBPN은 원본 이미지(I^l)를 컨볼루션 레이어들에 입력하여 특징 맵을 추출하고, 특징 맵을 업 블록 및 다운 블록이 반복적으로 구성된 모듈에 입력하는 신경망 모델일 수 있다. 업 블록은 특징 맵을 확대하는 블록일 수 있다. 다운 블록을 특징 맵을 축소시키는 블록일 수 있다. 반복적으로 구성된 업 블록 및 다운 블록은 각각 오차를 계산하여 네트워크에 피드백을 전달할 수 있다. 그리고, DBPN은 업 블록들 각각에서 생성되는 각각의 확대 특징 맵들(H^1~H^t)을 연결(concatenate)하는 신경망 모델일 수 있다. DBPN은 연결된 확대 특징 맵을 컨볼루션 레이어에 입력하여 고해상도 이미지인 제 1 변환 이미지(I^{sr})를 생성하는 신경망 모델일 수 있다.

[0084] 한편, 제 1 신경망 모델(200)은 기 설정된 제 1 손실 함수(loss function)에 기초하여 사전 학습될 수 있다. 제 1 손실 함수는 예측 값과 실제 값 간의 유사한 정도를 판단하는 기준일 수 있다. 제 1 손실 함수는 평균 제곱 오차(mean squared error), 평균 절대 오차(mean absolute error), 및/또는 평균 제곱근 오차(root mean square error) 중 적어도 하나를 포함할 수 있다.

[0085] 평균 제곱 오차는 예측 값과 실제 값 간의 평균을 제곱하여 평균을 낸 값일 수 있다.

[0086] 평균 절대 오차는 예측 값과 실제 값의 차이 값을 절댓값으로 변환하고, 변환된 차이 값들을 합산하며, 합산한 변환된 차이 값을 차이 값의 개수로 나누어 평균을 낸 값일 수 있다.

[0087] 평균 제곱근 오차는 평균 제곱 오차에 루트를 씌운 값일 수 있다. 따라서, 평균 제곱근 오차는 평균 제곱 오차에서 제곱을 통해 생기는 왜곡이 줄어들고, 오차를 보다 직관적으로 나타낼 수 있다.

[0088] 한편, 기 설정된 제 1 손실 함수는 다음 수학식 1로 표현될 수 있다.

수학식 1

$$(output_i - target_i)^2$$

[0089]

[0090] 여기서, i는 픽셀 번호(예를 들어, 300*300 해상도인 이미지의 경우, i는 1부터 90000까지의 숫자)이고, output은 제 1 변환 이미지의 각 픽셀이고, target은 원본 이미지 또는 제 1 적대적 이미지의 각 픽셀일 수 있다. 제 1 적대적 이미지는 제 1 적대적 이미지 생성부(500)에서 생성된 이미지일 수 있다. 제 1 적대적 이미지에 대한 자세한 설명은 제 1 적대적 이미지 생성부(500)을 설명하면 후술하도록 한다.

[0092] 다시 도 3을 참조하면, 제 2 신경망 모델(300)은 제 1 변환 이미지가 입력될 수 있고, 제 1 변환 이미지에 기초하여 제 1 물체 탐지 결과 데이터를 생성할 수 있다. 제 2 신경망 모델(300)은 제 1 변환 이미지에서 적어도 하나의 물체를 탐지하고, 그리고 적어도 하나의 물체의 위치 및 클래스(class)를 지정하여 제 1 물체 탐지 결과 데이터를 생성하는 제 1 물체 탐지 모델을 포함할 수 있다.

[0093] 제 2 신경망 모델(300)은 YOLO(You Only Look Once), Faster R-CNN(Region-based Convolutional Neural Networks) 및/또는 SSD(Single Shot Multibox Detector) 중 적어도 하나를 포함하는 제 1 물체 탐지 모델을 포

함할 수 있다.

[0094] 여기서, YOLO는 제 1 변환 이미지를 동일한 크기의 복수의 그리드들로 나누는 알고리즘일 수 있다. YOLO는 복수의 그리드들 각각에 대해 그리드 중앙을 중심으로 미리 정의된 형태로 지정된 바운딩박스(bounding-box)의 개수를 예측하는 알고리즘일 수 있다. YOLO는 바운딩박스의 개수에 기초하여 신뢰도를 계산하고, 이미지에 물체 포함 여부를 고려하여 높은 물체 신뢰도를 가지는 위치를 선택하여 물체를 파악하는 알고리즘일 수 있다. YOLO는 파악된 물체의 클래스를 지정하여 제 1 물체 탐지 결과 데이터를 생성할 수 있는 알고리즘일 수 있다.

[0095] Faster R-CNN은 제 1 변환 이미지에서 CNN을 통해 특징 맵을 추출하고, 특징 맵을 RPN(Region Proposal Net)을 이용하여 물체 영역후보들을 생성하는 신경망 모델일 수 있다. Faster R-CNN은 물체 영역후보들의 좌표 및 점수를 계산하고, 점수에 기초하여 물체의 클래스를 지정하여 제 1 물체 탐지 결과 데이터를 생성할 수 있는 신경망 모델일 수 있다.

[0096] SSD는 도 5를 참조하여 후술하도록 한다. 도 5는 본 개시의 몇몇 실시예에 따른 제 2 신경망 모델을 설명하기 위한 도면이다.

[0097] 도 5를 참조하면, SSD는 제 1 변환 이미지에서 CNN을 통해 특징 맵을 추출하는 신경망 모델일 수 있다. SSD는 서로 다른 복수의 컨볼루션 레이어들을 이용하여 다양한 크기의 복수의 특징 맵들을 추출하는 신경망 모델일 수 있다. SSD는 다양한 크기의 복수의 특징 맵들을 이용하여 물체 영역후보들의 좌표 및 클래스 별 점수를 계산하고, 클래스별 점수에 기초하여 물체의 클래스를 지정하여 제 1 물체 탐지 결과 데이터를 생성할 수 있는 신경망 모델일 수 있다.

[0098] 한편, 제 2 신경망 모델(300)은 기 설정된 제 2 손실 함수에 기초하여 사전 학습될 수 있다. 제 2 손실 함수는 예측 값과 실제 값 간의 유사한 정도를 판단하는 기준일 수 있다. 제 2 손실 함수는 평균 제곱 오차, 평균 절대 오차, 및/또는 평균 제곱근 오차 중 적어도 하나를 포함할 수 있다. 제 2 손실 함수는 Multi box loss(Multibox-loss)를 포함할 수 있다.

[0099] 제 2 손실 함수는 다음 수학식 2로 표현될 수 있다.

수학식 2

[0100]
$$L(x,c,l,g) = \frac{1}{N}(L_{conf}(x,c) + \alpha L_{loc}(x,l,g))$$

[0101] 여기서, L_{conf}는 제 2 신경망 모델(300)을 통해 생성된 제 1 물체 탐지 결과 데이터에 기초한 물체의 클래스와 기 저장된 실측 데이터에 기초한 실제 클래스에 대한 클래스 손실 값을 계산하는 수식이고, L_{loc}는 제 1 물체 탐지 결과 데이터에 기초한 물체의 위치와 기 저장된 실측 데이터에 기초한 실제 위치에 대한 위치 손실 값을 계산하는 수식이다.

[0103] 다시 도 3을 참조하면, 제 1 노이즈 생성부(400)는 제 1 물체 탐지 결과 데이터와 기 저장된 실측 데이터(ground-truth) 간의 제 1 손실(loss) 값에 기초하여 제 1 노이즈(noise)를 생성할 수 있다.

[0104] 제 1 노이즈 생성부(400)는 기 설정된 제 1 손실 함수에 기초하여 제 1 물체 탐지 결과 데이터와 기 저장된 실측 데이터 간의 제 1 손실 값을 산출할 수 있다. 제 1 노이즈 생성부(400)는 산출된 제 1 손실 값에 기초하여 제 1 노이즈를 생성할 수 있다.

[0106] 제 1 적대적 이미지 생성부(500)는 제 1 노이즈 및 제 1 변환 이미지에 기초하여 제 1 적대적 이미지를 생성할 수 있다. 제 1 적대적 이미지 생성부(500)는 제 1 변환 이미지를 구성하는 적어도 하나의 제 1 변환 이미지 픽셀에 제 1 노이즈를 추가하여 제 1 적대적 이미지를 생성할 수 있다. 즉, 제 1 적대적 이미지는 제 1 변환 이미지를 구성하는 적어도 하나의 제 1 변환 이미지 픽셀에 제 1 노이즈를 1회 이상 추가한 이미지일 수 있다.

[0107] 제 1 적대적 이미지 생성부(500)는 PGD(Projected Gradient Descent)를 포함할 수 있다. PGD는 다음 수학식 3으로 표현될 수 있다.

수학식 3

[0108] $$x^{t+1} = \Pi_{x+S}\left(x^t + \alpha\, sgn\left(\nabla_x L(\theta, x, y)\right)\right)$$

[0109] 여기서, x^t는 제 1 노이즈가 t번 추가된 제 1 적대적 이미지이고, $\nabla_x L(\theta, x, y)$는 제 1 노이즈이고, θ는 제 1 손실 함수의 파라미터(parameter)일 수 있다. 구체적으로, $\nabla_x L(\theta, x, y)$는 제 1 물체 탐지 결과 데이터와 기 저장된 실측 데이터 간의 차이를 역전파하여 산출된 x의 기울기인 제 1 노이즈를 의미할 수 있다. α는 노이즈의 강도를 설정하는 파라미터(parameter)일 수 있다.

[0111] 제 2 노이즈 생성부(600)는 제 1 적대적 이미지와 제 1 변환 이미지 간의 제 2 손실 값에 기초하여 제 2 노이즈를 생성할 수 있다.

[0112] 제 2 노이즈 생성부(600)는 기 설정된 제 2 손실 함수에 기초하여 제 1 적대적 이미지와 제 1 변환 이미지 간의 제 2 손실 값을 산출할 수 있다. 그리고, 제 2 노이즈 생성부(600)는 산출된 제 2 손실 값에 기초하여 제 2 노이즈를 생성할 수 있다.

[0114] 제 2 적대적 이미지 생성부(700)는 제 2 노이즈 및 원본 이미지에 기초하여 제 2 적대적 이미지를 생성할 수 있다. 제 2 적대적 이미지 생성부(700)는 원본 이미지를 구성하는 적어도 하나의 원본 이미지 픽셀에 제 2 노이즈를 추가하여 제 2 적대적 이미지를 생성할 수 있다. 즉, 제 2 적대적 이미지는 원본 이미지를 구성하는 적어도 하나의 원본 이미지 픽셀에 제 2 노이즈를 1회 이상 추가한 이미지일 수 있다.

[0115] 제 2 적대적 이미지 생성부(700)는 I_FGSM(Iterative Fast Gradient Signed Method)를 포함할 수 있다. I_FGSM은 FGSM을 반복적으로 수행하는 함수로 달리 말하면 I_FGSM의 최대 반복 횟수가 1인 함수는 FGSM과 동일할 수 있다. I_FGSM은 시간 소모가 크지만 일반적인 FGSM보다 공격효과가 더 뛰어날 수 있다.

[0116] FGSM은 신경망의 그래디언트(gradient)를 이용해 적대적 샘플을 생성하는 기법일 수 있다. 만약 모델의 입력이 이미지라면, 입력 이미지에 대한 손실 함수의 그래디언트를 계산하여 그 손실을 최대화하는 적대적 이미지를 생성할 수 있다.

[0117] I_FGSM은 다음 수학식 4로 표현될 수 있다.

수학식 4

[0118] $$\boldsymbol{X}_0^{adv} = \boldsymbol{X}, \quad \boldsymbol{X}_{N+1}^{adv} = Clip_{X,\epsilon}\left\{\boldsymbol{X}_N^{adv} + \alpha\, \mathrm{sign}\left(\nabla_X J(\boldsymbol{X}_N^{adv}, y_{true})\right)\right\}$$

[0119] 여기서 X_N^{adv}은 노이즈가 N번 추가된 이미지를 의미하며 N=0 인 경우, 즉 X_0^{adv}은 노이즈가 추가되지 않은 원본 이미지를 의미한다. y_{true}는 실측 데이터(ground-truth)를 의미하고, 본 실시예에서는 제 1 적대적 이미지일 수 있다. sign함수는 부호함수이다. $Clip_{X,\epsilon}$은 괄호 안의 값을 $X+\epsilon$과 $X-\epsilon$ 사이를 벗어나지 않게 클리핑(clipping) 해주어 노이즈가 사람의 눈에 잘 띄지 않게 해주는 역할을 할 수 있다. α는 노이즈의 강

도를 설정하는 파라미터(parameter)일 수 있다.

[0121] 제 3 신경망 모델(800)은 제 2 적대적 이미지에 기초하여 제 2 적대적 이미지보다 높은 해상도로 구성된 제 2 변환 이미지를 생성하는 제 2 초해상화 모델 및 제 2 변환 이미지에서 적어도 하나의 물체를 탐지하고, 그리고 탐지된 적어도 하나의 물체의 위치 및 클래스를 지정하여 제 2 물체 탐지 결과 데이터를 생성하는 제 2 물체 탐지 모델이 결합된 모델일 수 있다.

[0122] 구체적으로, 제 3 신경망 모델(800)은 SRCNN, RCAN 및/또는 DBPN 중 적어도 하나를 포함하는 제 2 초해상화 모델을 포함할 수 있다.

[0123] 여기서, SRCNN은 복수의 층(예를 들어, 세 개의 층)으로 구성된 신경망 모델일 수 있다. SRCNN은 제 2 적대적 이미지를 전처리 단계에서 바이큐빅(bicubic) 방식으로 다운 샘플링하여 저해상도 이미지를 생성하는 신경망 모델일 수 있다. 그리고, SRCNN은 저해상도 이미지에서부터 생성된 선명하지 않은 이미지와 제 2 적대적 이미지 사이의 특징을 학습하는 방식으로 사전 학습될 수 있다. SRCNN은 입력된 제 2 적대적 이미지를 제 2 적대적 이미지보다 높은 해상도로 구성된 제 2 변환 이미지를 생성하는 신경망 모델일 수 있다.

[0124] RCAN은 컨볼루션 레이어를 통해 제 2 적대적 이미지의 특징 정보를 추출하는 신경망 모델일 수 있다. 그리고, RCAN은 channel attention module을 통해 특징 정보에서 심층 특징 정보를 추출하여 고해상도 이미지인 제 2 변환 이미지를 생성하는 신경망 모델일 수 있다.

[0125] DBPN은 제 2 적대적 이미지를 컨볼루션 레이어들에 입력하여 특징 맵을 추출하고, 특징 맵을 업 블록 및 다운 블록이 반복적으로 구성된 모듈에 입력하는 신경망 모델일 수 있다. 업 블록은 특징 맵을 확대하는 블록일 수 있다. 다운 블록을 특징 맵을 축소시키는 블록일 수 있다. 반복적으로 구성된 업 블록 및 다운 블록은 각각 오차를 계산하여 네트워크에 피드백을 전달할 수 있다. 그리고, DBPN은 업 블록들 각각에서 생성되는 각각의 확대 특징 맵들을 연결(concatenate)하는 신경망 모델일 수 있다. DBPN은 연결된 확대 특징 맵을 컨볼루션 레이어에 입력하여 고해상도 이미지인 제 2 변환 이미지를 생성하는 신경망 모델일 수 있다.

[0127] 한편, 제 3 신경망 모델(800)은 YOLO, Faster R-CNN 및/또는 SSD중 적어도 하나를 포함하는 제 2 물체 탐지 모델을 포함할 수 있다.

[0128] 여기서, YOLO는 제 2 변환 이미지를 동일한 크기의 복수의 그리드들로 나누는 알고리즘일 수 있다. YOLO는 복수의 그리드들 각각에 대해 그리드 중앙을 중심으로 미리 정의된 형태로 지정된 바운딩박스(bounding-box)의 개수를 예측하는 알고리즘일 수 있다. YOLO는 바운딩박스의 개수에 기초하여 신뢰도를 계산하고, 이미지에 물체 포함 여부를 고려하여 높은 물체 신뢰도를 가지는 위치를 선택하여 물체를 파악하는 알고리즘일 수 있다. YOLO는 파악된 물체의 클래스를 지정하여 제 2 물체 탐지 결과 데이터를 생성할 수 있는 알고리즘일 수 있다.

[0129] Faster R-CNN은 제 2 변환 이미지에서 CNN을 통해 특징 맵을 추출하고, 특징 맵을 RPN(Region Proposal Net)을 이용하여 물체 영역후보들을 생성하는 신경망 모델일 수 있다. Faster R-CNN은 물체 영역후보들의 좌표 및 점수를 계산하고, 점수에 기초하여 물체의 클래스를 지정하여 제 2 물체 탐지 결과 데이터를 생성할 수 있는 신경망 모델일 수 있다.

[0130] SSD는 제 2 변환 이미지에서 CNN을 통해 특징 맵을 추출하는 신경망 모델일 수 있다. SSD는 서로 다른 복수의 컨볼루션 레이어들을 이용하여 다양한 크기의 복수의 특징 맵들을 추출하는 신경망 모델일 수 있다. SSD는 다양한 크기의 복수의 특징 맵들을 이용하여 물체 영역후보들의 좌표 및 클래스 별 점수를 계산하고, 클래스별 점수에 기초하여 물체의 클래스를 지정하여 제 2 물체 탐지 결과 데이터를 생성할 수 있는 신경망 모델일 수 있다.

[0131] 한편, 본 개시의 몇몇 실시예에 따른 제 3 신경망 모델(800)은 도 6을 참조하여 후술하도록 한다. 도 6은 본 개시의 몇몇 실시예에 따른 제 3 신경망 모델(800)을 설명하기 위한 도면이다.

[0132] 도 6을 참조하면, 제 3 신경망 모델(800)은 DBPN을 포함하는 제 2 초해상화 모델과 SSD를 포함하는 제 2 물체 탐지 모델이 결합된 모델일 수 있다. 따라서, 제 3 신경망 모델(800)은 제 2 적대적 이미지(X^l)를 DBPN을 포함하는 제 2 초해상화 모델에 입력하여 고해상도 이미지인 제 2 변환 이미지(X^h)를 생성할 수 있다. 그리고, 제 3 신경망 모델(800)은 생성된 제 2 변환 이미지(X^h)를 SSD를 포함하는 제 2 물체 탐지 모델에 입력하여 제 2 물

체 탐지 결과 데이터($\hat{y}(X^h)$)를 생성할 수 있다.

[0134] 제 3 신경망 모델(800)은 제 2 적대적 이미지가 입력될 수 있고, 제 2 적대적 이미지를 통해 제 3 신경망 모델(800)의 성능이 판단될 수 있다. 제 3 신경망 모델(800)은 제 2 적대적 이미지에 기초하여 제 2 물체 탐지 결과 데이터를 생성할 수 있다.

[0135] 한편, 프로세서(110)는 제 2 물체 탐지 결과 데이터와 기 저장된 실측 데이터 간의 제 3 손실 값에 기초하여 제 3 신경망 모델(800)의 성능을 판단할 수 있다. 프로세서(110)는 제 3 손실 값과 기 설정된 임계 값의 비교를 통해 제 3 신경망 모델(800)의 성능을 판단할 수 있다. 예를 들어, 프로세서(110)는 제 3 손실 값이 기 설정된 임계 값보다 큰 경우, 제 3 신경망 모델(800)의 성능을 낮게 판단하여 기준 미달로 판단할 수 있다. 다만, 프로세서(110)에서 제 3 신경망 모델(800)의 성능을 판단하는 방법은 이에 제한되지 않으며, 프로세서(110)는 다양한 방법을 통해 제 3 신경망 모델(800)의 성능을 판단할 수 있다.

[0136] 여기서, 본 개시의 몇몇 실시예에 따르면, 제 3 신경망 모델은 외부 기기에 존재할 수 있다. 따라서, 프로세서(110)는 제 2 적대적 이미지를 네트워크부(150)를 통해 외부 기기로 전달할 수 있다. 외부 기기는 제 3 신경망 모델에 수신한 제 2 적대적 이미지를 입력하여 제 3 신경망 모델의 성능을 판단하고, 제 3 신경망 모델을 보완할 수 있다.

[0137] 한편, 본 개시의 몇몇 실시예에 따르면, 제 1 신경망 모델(200) 및 제 2 신경망 모델(300)이 서로 결합되어 하나의 신경망 모델로 구성될 수도 있다. 따라서, 제 1 신경망 모델(200) 및 제 2 신경망 모델(300)이 서로 결합된 하나의 모델로 구성되어 제 3 신경망 모델(800)과 유사한 구조를 가질 수도 있다.

[0139] 도 7은 본 개시의 몇몇 실시예에 따른 컴퓨팅 장치에 의해 수행되는 적대적 공격을 수행하는 방법을 설명하기 위한 순서도이다.

[0140] 도 7을 참조하면, 컴퓨팅 장치(100)의 프로세서(110)는 원본 이미지를 제 1 신경망 모델(200)에 입력하여 제 1 변환 이미지를 생성할 수 있다(S110).

[0141] 여기서, 제 1 신경망 모델(200)은 원본 이미지에 기초하여 원본 이미지보다 높은 해상도로 구성된 제 1 변환 이미지를 생성하는 제 1 초해상화 모델을 포함할 수 있다.

[0142] 또한, 제 1 신경망 모델(200)은 기 설정된 제 1 손실 함수에 기초하여 사전 학습될 수 있다.

[0144] 프로세서(110)는 제 1 변환 이미지를 제 2 신경망 모델(300)에 입력하여 제 1 물체 탐지 결과 데이터를 생성할 수 있다(S120).

[0145] 여기서, 제 2 신경망 모델(300)은 제 1 변환 이미지에서 적어도 하나의 물체를 탐지하고, 그리고 적어도 하나의 물체의 위치 및 클래스를 지정하여 제 1 물체 탐지 결과 데이터를 생성하는 제 1 물체 탐지 모델을 포함할 수 있다.

[0147] 프로세서(110)는 제 1 물체 탐지 결과 데이터와 기 저장된 실측 데이터(ground-truth) 간의 제 1 손실(loss) 값에 기초하여 제 1 노이즈(noise)를 생성할 수 있다(S130).

[0148] 프로세서(110)는 기 설정된 제 1 손실 함수에 기초하여 제 1 물체 탐지 결과 데이터와 기 저장된 실측 데이터 간의 제 1 손실 값을 산출할 수 있다. 프로세서(110)는 산출된 제 1 손실 값에 기초하여 제 1 노이즈를 생성할 수 있다.

[0150] 프로세서(110)는 제 1 노이즈 및 제 1 변환 이미지에 기초하여 제 1 적대적 이미지를 생성할 수 있다(S140).

[0151] 프로세서(110)는 제 1 변환 이미지를 구성하는 적어도 하나의 제 1 변환 이미지 픽셀에 제 1 노이즈를 추가하여 제 1 적대적 이미지를 생성할 수 있다.

[0153] 프로세서(110)는 제 1 적대적 이미지와 제 1 변환 이미지 간의 제 2 손실 값에 기초하여 제 2 노이즈를 생성할 수 있다(S150).

[0154] 프로세서(110)는 기 설정된 제 2 손실 함수에 기초하여 제 1 적대적 이미지와 제 1 변환 이미지 간의 제 2 손실 값을 산출할 수 있다. 프로세서(110)는 산출된 제 2 손실 값에 기초하여 제 2 노이즈를 생성할 수 있다.

[0156] 프로세서(110)는 제 2 노이즈 및 원본 이미지에 기초하여 제 2 적대적 이미지를 생성할 수 있다(S160).

[0157] 프로세서(110)는 원본 이미지를 구성하는 적어도 하나의 원본 이미지 픽셀에 제 2 노이즈를 추가하여 제 2 적대적 이미지를 생성할 수 있다.

[0159] 프로세서(110)는 제 2 적대적 이미지를 제 3 신경망 모델(800)에 입력하여 제 3 신경망 모델(800)의 성능을 판단할 수 있다(S170).

[0160] 프로세서(110)는 제 2 적대적 이미지를 제 3 신경망 모델에 입력하여 제 2 물체 탐지 결과 데이터를 생성할 수 있다. 프로세서(110)는 제 2 물체 탐지 결과 데이터와 기 저장된 실측 데이터 간의 제 3 손실 값에 기초하여 제 3 신경망 모델의 성능을 판단할 수 있다.

[0161] 여기서, 제 3 신경망 모델(800)은 제 2 적대적 이미지에 기초하여 제 2 적대적 이미지보다 높은 해상도로 구성된 제 2 변환 이미지를 생성하는 제 2 초해상화 모델 및 제 2 변환 이미지에서 적어도 하나의 물체를 탐지하고, 그리고 탐지된 적어도 하나의 물체의 위치 및 클래스를 지정하여 제 2 물체 탐지 결과 데이터를 생성하는 제 2 물체 탐지 모델이 결합된 모델일 수 있다.

[0163] 도 7에 도시되는 단계들은 예시적인 단계들이다. 따라서, 본 개시내용의 사상의 범위를 벗어나지 않는 한도에서 도 7의 단계들 중 일부가 생략되거나 추가적인 단계들이 존재할 수 있다는 점 또한 당업자에게 명백할 것이다.

[0164] 또한, 도 7에 기재된 구성들(100 내지 800) 및 각 단계에 관한 구체적인 내용은 앞서 도 1 내지 도 6을 통해 설명한 내용으로 대체될 수 있다.

[0166] 한편, 도 1 내지 도 7을 참조하여 상술한 컴퓨팅 장치(100)에 의해 적대적 이미지를 생성하는 방법 중에서 하나의 실시예를 후술하도록 한다.

[0168] [실시예]

[0169] 컴퓨팅 장치(100)의 프로세서(110)는 75*75 해상도의 이미지를 원본 이미지로 지정하였다.

[0170] 1단계: 프로세서(110)는 DBPN을 포함하는 제 1 신경망 모델(200)에 원본 이미지를 입력하여 해상도를 4배 증가시킨 300*300 해상도의 제 1 변환 이미지를 생성하였다.

[0171] 2단계: 프로세서(110)는 SSD를 포함하는 제 2 신경망 모델(300)에 제 1 변환 이미지를 입력하여 제 1 물체 탐지 결과 데이터를 생성하고, 제 1 물체 탐지 결과 데이터와 기 저장된 실측 데이터 간의 제 1 손실 값을 제 2 손실 함수를 이용하여 산출하였다.

[0172] 3단계: 프로세서(110)는 제 1 손실 값을 역전파(backward)하여 제 1 변환 이미지의 기울기를 산출하였다.

[0173] 4단계: 프로세서(110)는 상술한 수학식 3에서 $\nabla_x L(\theta, x, y)$ 부분에 제 1 변환 이미지의 기울기를 대입하고 부호 함수를 통해 제 1 노이즈의 부호를 계산하였다.

[0174] 5단계: 프로세서(110)는 제 1 노이즈의 부호에 노이즈 강도 값인 α 와 곱하였다.

[0175] 6단계: 프로세서(110)는 5단계에서 산출된 값을 제 1 변환 이미지와 더하였다.

[0176] 7단계: 프로세서(110)는 6단계에서 산출된 값에서 제 1 변환 이미지를 뺐다.

[0177] 8단계: 프로세서(110)는 7단계에서 산출된 값을 $\pm\epsilon$ 사이로 클리핑(clipping)하여 제 1 노이즈를 산출하였다.

[0178] 9단계: 프로세서(110)는 제 1 변환 이미지에 제 1 노이즈를 더하여 노이즈 생성이 1번 반복된 1차 제 1 적대적 이미지를 산출하였다.

[0179] 10단계: 프로세서(110)는 2단계부터 9단계까지 반복하되, 2, 3, 6단계의 제 1 변환 이미지를 1차 제 1 적대적 이미지로 대체하여 진행하였다. 프로세서(110)는 이를 최대 반복 횟수만큼 반복하며 다음 번 반복에서 1차 제 1 적대적 이미지를 2차 제 1 적대적 이미지로 대체하여 진행하였다.

[0180] 즉, 프로세서(110)는 (N-1)차 제 1 적대적 이미지를 N차 제 1 적대적 이미지로 대체하여 2단계부터 9단계까지 반복하고, 최종적으로 최대 반복 횟수인 N번 반복한 뒤 나온 N차 제 1 적대적 이미지를 최종 제 1 적대적 이미지, 즉, 본 개시에 따른 제 1 적대적 이미지로 결정하였다.

[0181] 11단계: 프로세서(110)는 제 1 변환 이미지와 제 1 적대적 이미지 간의 제 2 손실 값을 제 1 손실 함수를 이용하여 산출하였다.

[0182] 12단계: 프로세서(110)는 제 2 손실 값을 역전파하여 원본 이미지의 기울기를 산출하였다.

[0183] 13단계: 프로세서(110)는 수학식 4에서 $\nabla_X J(\boldsymbol{X}_N^{adv}, y_{true})$ 부분에 원본 이미지의 기울기를 대입하여 제 2 노이즈의 부호를 계산하였다.

[0184] 14단계: 프로세서(110)는 제 2 노이즈 부호에 α 를 최대 반복 횟수인 N으로 나눈 값을 곱하였다.

[0185] 15단계: 프로세서(110)는 14단계에서 산출한 값을 $\pm\epsilon$ 사이로 클리핑(clipping)하여 제 2 노이즈를 산출하였다.

[0186] 16단계: 프로세서(110)는 원본 이미지와 제 2 노이즈를 더하고, 이를 0에서 1사이로 클리핑(clipping)하여 1차 제 2 적대적 이미지를 산출한다.

[0187] 17단계: 프로세서(110)는 11단계부터 16단계까지 반복하되, 11단계의 제 1 변환 이미지를 1차 제 2 적대적 이미지를 초해상화한 1차 제 2 변환 적대적 이미지로 대체하고, 12, 16단계의 원본 이미지를 1차 제 2 적대적 이미지로 대체하여 진행하였다.

[0188] 즉, 프로세서(110)는 (N-1)차 제 2 적대적 이미지를 N차 제 2 적대적 이미지로 대체하여 11단계부터 16단계까지 반복하고, 최종적으로 최대 반복 횟수인 N번 반복한 뒤 나온 N차 제 2 적대적 이미지를 최종 제 2 적대적 이미지, 즉, 본 개시에 따른 제 2 적대적 이미지로 결정하였다.

[0189] 따라서, 프로세서(110)는 초해상화 모델을 포함하는 제 1 신경망 모델의 학습에 사용되는 제 1 손실 함수를 이용하여 생성한 노이즈 및 물체 탐지 모델을 포함하는 제 2 신경망 모델의 학습에 사용되는 제 2 손실 함수를 이용하여 생성한 노이즈를 모두 포함하도록 제 2 적대적 이미지를 생성할 수 있다. 즉, 프로세서(110)는 초해상화 모델과 물체 탐지 모델에 대해 모두 공격을 수행할 수 있는 제 2 적대적 이미지를 생성할 수 있다.

[0190] 프로세서(110)는 제 2 적대적 이미지를 이용하여 초해상화가 탑재된 물체 탐지 모델의 오작동을 유도하는 적대적 공격을 수행할 수 있다.

[0191] 상술한 바에 따르면, 본 개시는 기존의 하나의 모델에 대해서만 공격을 수행하는 개별적 공격과 달리 2가지 모델에 대해서 공격을 수행하여 높은 성능을 가질 수 있다. 본 개시는 초해상화가 탑재된 물체 탐지 모델에 적대적 공격을 수행하여 해당 모델의 평가를 통해 견고성(robustness)을 측정할 수 있다. 본 개시는 컴퓨터 비전 모델의 방어 방법에 대한 연구의 시야를 넓힐 수 있다.

[0192] 적대적 공격은 특정 모델에 대해 수행되는데, 이에 대한 방어 기법 역시 특정 적대적 공격에 대한 대해서만 뛰어난 방어 성능을 보이게 된다. 그러나 두 모델이 조합되어 사용되는 환경에서 하나의 모델에 대해서만 개별적 방어책을 구성할 경우, 두 모델을 모두 공격하는 적대적 이미지에 대해서는 방어 성능이 저하된다. 따라서, 본 개시의 일 실시예에 따른 기법에 따르면, 두 모델이 조합되어 사용되는 환경에서 두 모델에 대한 평가를 위해 두 모델을 모두 공격할 수 있는 적대적 이미지의 생성이 가능해질 수 있다.

[0194] 도 8은 본 개시의 실시예들이 구현될 수 있는 예시적인 컴퓨팅 환경에 대한 간략하고 일반적인 개략도를 도시한다.

[0195] 본 개시가 일반적으로 컴퓨팅 장치에 의해 구현될 수 있는 것으로 전술되었지만, 당업자라면 본 개시가 하나 이상의 컴퓨터 상에서 실행될 수 있는 컴퓨터 실행가능 명령어 및/또는 기타 프로그램 모듈들과 결합되어 및/또는 하드웨어와 소프트웨어의 조합으로써 구현될 수 있다는 것을 잘 알 것이다.

[0196] 일반적으로, 프로그램 모듈은 특정의 태스크를 수행하거나 특정의 추상 데이터 유형을 구현하는 루틴, 프로그램, 컴포넌트, 데이터 구조, 기타 등등을 포함한다. 또한, 당업자라면 본 개시의 방법이 단일-프로세서 또는 멀티프로세서 컴퓨터 시스템, 미니컴퓨터, 메인프레임 컴퓨터는 물론 퍼스널 컴퓨터, 핸드헬드(handheld) 컴퓨팅 장치, 마이크로프로세서-기반 또는 프로그램가능 가전 제품, 기타 등등(이들 각각은 하나 이상의 연관된 장치와 연결되어 동작할 수 있음)을 비롯한 다른 컴퓨터 시스템 구성으로 실시될 수 있다는 것을 잘 알 것이다.

[0197] 본 개시의 설명된 실시예들은 또한 어떤 태스크들이 통신 네트워크를 통해 연결되어 있는 원격 처리 장치들에 의해 수행되는 분산 컴퓨팅 환경에서 실시될 수 있다. 분산 컴퓨팅 환경에서, 프로그램 모듈은 로컬 및 원격 메모리 저장 장치 둘 다에 위치할 수 있다.

[0198] 컴퓨터는 통상적으로 다양한 컴퓨터 판독가능 매체를 포함한다. 컴퓨터에 의해 액세스 가능한 매체는 그 어떤 것이든지 컴퓨터 판독가능 매체가 될 수 있고, 이러한 컴퓨터 판독가능 매체는 휘발성 및 비휘발성 매체, 일시적(transitory) 및 비일시적(non-transitory) 매체, 이동식 및 비-이동식 매체를 포함한다. 제한이 아닌 예로서, 컴퓨터 판독가능 매체는 컴퓨터 판독가능 저장 매체 및 컴퓨터 판독가능 전송 매체를 포함할 수 있다. 컴퓨터 판독가능 저장 매체는 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈 또는 기타 데이터와 같은 정보를 저장하는 임의의 방법 또는 기술로 구현되는 휘발성 및 비휘발성 매체, 일시적 및 비-일시적 매체, 이동식 및 비이동식 매체를 포함한다. 컴퓨터 판독가능 저장 매체는 RAM, ROM, EEPROM, 플래시 메모리 또는 기타 메모리 기술, CD-ROM, DVD(digital video disk) 또는 기타 광 디스크 저장 장치, 자기 카세트, 자기 테이프, 자기 디스크 저장 장치 또는 기타 자기 저장 장치, 또는 컴퓨터에 의해 액세스될 수 있고 원하는 정보를 저장하는 데 사용될 수 있는 임의의 기타 매체를 포함하지만, 이에 한정되지 않는다.

[0199] 컴퓨터 판독가능 전송 매체는 통상적으로 반송파(carrier wave) 또는 기타 전송 메커니즘(transport mechanism)과 같은 피변조 데이터 신호(modulated data signal)에 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈 또는 기타 데이터 등을 구현하고 모든 정보 전달 매체를 포함한다. 피변조 데이터 신호라는 용어는 신호 내에 정보를 인코딩하도록 그 신호의 특성들 중 하나 이상을 설정 또는 변경시킨 신호를 의미한다. 제한이 아닌 예로서, 컴퓨터 판독가능 전송 매체는 유선 네트워크 또는 직접 배선 접속(direct-wired connection)과 같은 유선 매체, 그리고 음향, RF, 적외선, 기타 무선 매체와 같은 무선 매체를 포함한다. 상술된 매체들 중 임의의 것의 조합도 역시 컴퓨터 판독가능 전송 매체의 범위 안에 포함되는 것으로 한다.

[0200] 컴퓨터(1102)를 포함하는 본 개시의 여러가지 측면들을 구현하는 예시적인 환경(1100)이 나타내어져 있으며, 컴퓨터(1102)는 처리 장치(1104), 시스템 메모리(1106) 및 시스템 버스(1108)를 포함한다. 시스템 버스(1108)는 시스템 메모리(1106)(이에 한정되지 않음)를 비롯한 시스템 컴포넌트들을 처리 장치(1104)에 연결시킨다. 처리 장치(1104)는 다양한 상용 프로세서들 중 임의의 프로세서일 수 있다. 듀얼 프로세서 및 기타 멀티프로세서 아키텍처도 역시 처리 장치(1104)로서 이용될 수 있다.

[0201] 시스템 버스(1108)는 메모리 버스, 주변장치 버스, 및 다양한 상용 버스 아키텍처 중 임의의 것을 사용하는 로컬 버스에 추가적으로 상호 연결될 수 있는 몇 가지 유형의 버스 구조 중 임의의 것일 수 있다. 시스템 메모리(1106)는 판독 전용 메모리(ROM)(1110) 및 랜덤 액세스 메모리(RAM)(1112)를 포함한다. 기본 입/출력 시스템(BIOS)은 ROM, EPROM, EEPROM 등의 비휘발성 메모리(1110)에 저장되며, 이 BIOS는 시동 중과 같은 때에 컴퓨터(1102) 내의 구성요소들 간에 정보를 전송하는 일을 돕는 기본적인 루틴을 포함한다. RAM(1112)은 또한 데이터를 캐싱하기 위한 정적 RAM 등의 고속 RAM을 포함할 수 있다.

[0202] 컴퓨터(1102)는 또한 내장형 하드 디스크 드라이브(HDD)(1114)(예를 들어, EIDE, SATA)-이 내장형 하드 디스크 드라이브(1114)는 또한 적당한 섀시(도시 생략) 내에서 외장형 용도로 구성될 수 있음-, 자기 플로피 디스크 드라이브(FDD)(1116)(예를 들어, 이동식 디스켓(1118)으로부터 판독을 하거나 그에 기록을 하기 위한 것임), 및 광 디스크 드라이브(1120)(예를 들어, CD-ROM 디스크(1122)를 판독하거나 DVD 등의 기타 고용량 광 매체로부터 판독을 하거나 그에 기록을 하기 위한 것임)를 포함한다. 하드 디스크 드라이브(1114), 자기 디스크 드라이브

(1116) 및 광 디스크 드라이브(1120)는 각각 하드 디스크 드라이브 인터페이스(1124), 자기 디스크 드라이브 인터페이스(1126) 및 광 드라이브 인터페이스(1128)에 의해 시스템 버스(1108)에 연결될 수 있다. 외장형 드라이브 구현을 위한 인터페이스(1124)는 USB(Universal Serial Bus) 및 IEEE 1394 인터페이스 기술 중 적어도 하나 또는 그 둘 다를 포함한다.

[0203] 이들 드라이브 및 그와 연관된 컴퓨터 판독가능 매체는 데이터, 데이터 구조, 컴퓨터 실행가능 명령어, 기타 등등의 비휘발성 저장을 제공한다. 컴퓨터(1102)의 경우, 드라이브 및 매체는 임의의 데이터를 적당한 디지털 형식으로 저장하는 것에 대응한다. 상기에서의 컴퓨터 판독가능 매체에 대한 설명이 HDD, 이동식 자기 디스크, 및 CD 또는 DVD 등의 이동식 광 매체를 언급하고 있지만, 당업자라면 집 드라이브(zip drive), 자기 카세트, 플래쉬 메모리 카드, 카트리지, 기타 등등의 컴퓨터에 의해 판독가능한 다른 유형의 매체도 역시 예시적인 운영 환경에서 사용될 수 있으며 또 임의의 이러한 매체가 본 개시의 방법들을 수행하기 위한 컴퓨터 실행가능 명령어를 포함할 수 있다는 것을 잘 알 것이다.

[0204] 운영 체제(1130), 하나 이상의 애플리케이션 프로그램(1132), 기타 프로그램 모듈(1134) 및 프로그램 데이터(1136)를 비롯한 다수의 프로그램 모듈이 드라이브 및 RAM(1112)에 저장될 수 있다. 운영 체제, 애플리케이션, 모듈 및/또는 데이터의 전부 또는 그 일부분이 또한 RAM(1112)에 캐싱될 수 있다. 본 개시가 여러가지 상업적으로 이용가능한 운영 체제 또는 운영 체제들의 조합에서 구현될 수 있다는 것을 잘 알 것이다.

[0205] 사용자는 하나 이상의 유선/무선 입력 장치, 예를 들어, 키보드(1138) 및 마우스(1140) 등의 포인팅 장치를 통해 컴퓨터(1102)에 명령 및 정보를 입력할 수 있다. 기타 입력 장치(도시 생략)로는 마이크, IR 리모콘, 조이스틱, 게임 패드, 스타일러스 펜, 터치 스크린, 기타 등등이 있을 수 있다. 이들 및 기타 입력 장치가 종종 시스템 버스(1108)에 연결되어 있는 입력 장치 인터페이스(1142)를 통해 처리 장치(1104)에 연결되지만, 병렬 포트, IEEE 1394 직렬 포트, 게임 포트, USB 포트, IR 인터페이스, 기타 등등의 기타 인터페이스에 의해 연결될 수 있다.

[0206] 모니터(1144) 또는 다른 유형의 디스플레이 장치도 역시 비디오 어댑터(1146) 등의 인터페이스를 통해 시스템 버스(1108)에 연결된다. 모니터(1144)에 부가하여, 컴퓨터는 일반적으로 스피커, 프린터, 기타 등등의 기타 주변 출력 장치(도시 생략)를 포함한다.

[0207] 컴퓨터(1102)는 유선 및/또는 무선 통신을 통한 원격 컴퓨터(들)(1148) 등의 하나 이상의 원격 컴퓨터로의 논리적 연결을 사용하여 네트워크화된 환경에서 동작할 수 있다. 원격 컴퓨터(들)(1148)는 워크스테이션, 컴퓨팅 디바이스 컴퓨터, 라우터, 퍼스널 컴퓨터, 휴대용 컴퓨터, 마이크로프로세서-기반 오락 기기, 피어 장치 또는 기타 통상의 네트워크 노드일 수 있으며, 일반적으로 컴퓨터(1102)에 대해 기술된 구성요소들 중 다수 또는 그 전부를 포함하지만, 간략함을 위해, 메모리 저장 장치(1150)만이 도시되어 있다. 도시되어 있는 논리적 연결은 근거리 통신망(LAN)(1152) 및/또는 더 큰 네트워크, 예를 들어, 원거리 통신망(WAN)(1154)에의 유선/무선 연결을 포함한다. 이러한 LAN 및 WAN 네트워킹 환경은 사무실 및 회사에서 일반적인 것이며, 인트라넷 등의 전사적 컴퓨터 네트워크(enterprise-wide computer network)를 용이하게 해주며, 이들 모두는 전세계 컴퓨터 네트워크, 예를 들어, 인터넷에 연결될 수 있다.

[0208] LAN 네트워킹 환경에서 사용될 때, 컴퓨터(1102)는 유선 및/또는 무선 통신 네트워크 인터페이스 또는 어댑터(1156)를 통해 로컬 네트워크(1152)에 연결된다. 어댑터(1156)는 LAN(1152)에의 유선 또는 무선 통신을 용이하게 해줄 수 있으며, 이 LAN(1152)은 또한 무선 어댑터(1156)와 통신하기 위해 그에 설치되어 있는 무선 액세스 포인트를 포함하고 있다. WAN 네트워킹 환경에서 사용될 때, 컴퓨터(1102)는 모뎀(1158)을 포함할 수 있거나, WAN(1154) 상의 통신 컴퓨팅 디바이스에 연결되거나, 또는 인터넷을 통하는 등, WAN(1154)을 통해 통신을 설정하는 기타 수단을 갖는다. 내장형 또는 외장형 및 유선 또는 무선 장치일 수 있는 모뎀(1158)은 직렬 포트 인터페이스(1142)를 통해 시스템 버스(1108)에 연결된다. 네트워크화된 환경에서, 컴퓨터(1102)에 대해 설명된 프로그램 모듈들 또는 그의 일부분이 원격 메모리/저장 장치(1150)에 저장될 수 있다. 도시된 네트워크 연결이 예시적인 것이며 컴퓨터들 사이에 통신 링크를 설정하는 기타 수단이 사용될 수 있다는 것을 잘 알 것이다.

[0209] 컴퓨터(1102)는 무선 통신으로 배치되어 동작하는 임의의 무선 장치 또는 개체, 예를 들어, 프린터, 스캐너, 데스크톱 및/또는 휴대용 컴퓨터, PDA(portable data assistant), 통신 위성, 무선 검출가능 태그와 연관된 임의의 장비 또는 장소, 및 전화와 통신을 하는 동작을 한다. 이것은 적어도 Wi-Fi 및 블루투스 무선 기술을 포함한다. 따라서, 통신은 종래의 네트워크에서와 같이 미리 정의된 구조이거나 단순하게 적어도 2개의 장치 사이의 애드혹 통신(ad hoc communication)일 수 있다.

[0210] Wi-Fi(Wireless Fidelity)는 유선 없이도 인터넷 등으로의 연결을 가능하게 해준다. Wi-Fi는 이러한 장치, 예를 들어, 컴퓨터가 실내에서 및 실외에서, 즉 기지국의 통화권 내의 아무 곳에서나 데이터를 전송 및 수신할 수 있게 해주는 셀 전화와 같은 무선 기술이다. Wi-Fi 네트워크는 안전하고 신뢰성 있으며 고속인 무선 연결을 제공하기 위해 IEEE 802.11(a, b, g, 기타)이라고 하는 무선 기술을 사용한다. 컴퓨터를 서로에, 인터넷에 및 유선 네트워크(IEEE 802.3 또는 이더넷을 사용함)에 연결시키기 위해 Wi-Fi가 사용될 수 있다. Wi-Fi 네트워크는 비인가 2.4 및 5GHz 무선 대역에서, 예를 들어, 11Mbps(802.11a) 또는 54 Mbps(802.11b) 데이터 레이트로 동작하거나, 양 대역(듀얼 대역)을 포함하는 제품에서 동작할 수 있다.

[0211] 본 개시의 기술 분야에서 통상의 지식을 가진 자는 정보 및 신호들이 임의의 다양한 상이한 기술들 및 기법들을 이용하여 표현될 수 있다는 것을 이해할 것이다. 예를 들어, 위의 설명에서 참조될 수 있는 데이터, 지시들, 명령들, 정보, 신호들, 비트들, 심볼들 및 칩들은 전압들, 전류들, 전자기파들, 자기장들 또는 입자들, 광학장들 또는 입자들, 또는 이들의 임의의 결합에 의해 표현될 수 있다.

[0212] 본 개시의 기술 분야에서 통상의 지식을 가진 자는 여기에 개시된 실시예들과 관련하여 설명된 다양한 예시적인 논리 블록들, 모듈들, 프로세서들, 수단들, 회로들 및 알고리즘 단계들이 전자 하드웨어, (편의를 위해, 여기에서 소프트웨어로 지칭되는) 다양한 형태들의 프로그램 또는 설계 코드 또는 이들 모두의 결합에 의해 구현될 수 있다는 것을 이해할 것이다. 하드웨어 및 소프트웨어의 이러한 상호 호환성을 명확하게 설명하기 위해, 다양한 예시적인 컴포넌트들, 블록들, 모듈들, 회로들 및 단계들이 이들의 기능과 관련하여 위에서 일반적으로 설명되었다. 이러한 기능이 하드웨어 또는 소프트웨어로서 구현되는지 여부는 특정한 애플리케이션 및 전체 시스템에 대하여 부과되는 설계 제약들에 따라 좌우된다. 본 개시의 기술 분야에서 통상의 지식을 가진 자는 각각의 특정한 애플리케이션에 대하여 다양한 방식들로 설명된 기능을 구현할 수 있으나, 이러한 구현 결정들은 본 개시의 범위를 벗어나는 것으로 해석되어서는 안 될 것이다.

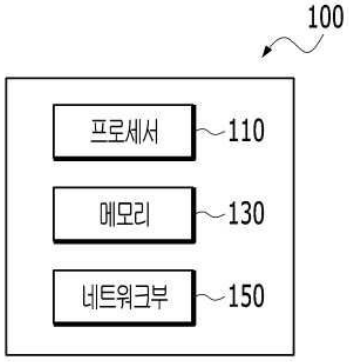
[0213] 여기서 제시된 다양한 실시예들은 방법, 장치, 또는 표준 프로그래밍 및/또는 엔지니어링 기술을 사용한 제조 물품(article)으로 구현될 수 있다. 용어 제조 물품은 임의의 컴퓨터-판독가능 저장장치로부터 액세스 가능한 컴퓨터 프로그램, 캐리어, 또는 매체(media)를 포함한다. 예를 들어, 컴퓨터-판독가능 저장매체는 자기 저장 장치(예를 들면, 하드 디스크, 플로피 디스크, 자기 스트립, 등), 광학 디스크(예를 들면, CD, DVD, 등), 스마트 카드, 및 플래쉬 메모리 장치(예를 들면, EEPROM, 카드, 스틱, 키 드라이브, 등)를 포함하지만, 이들로 제한되는 것은 아니다. 또한, 여기서 제시되는 다양한 저장 매체는 정보를 저장하기 위한 하나 이상의 장치 및/또는 다른 기계-판독가능한 매체를 포함한다.

[0214] 제시된 프로세스들에 있는 단계들의 특정한 순서 또는 계층 구조는 예시적인 접근들의 일례임을 이해하도록 한다. 설계 우선순위들에 기반하여, 본 개시의 범위 내에서 프로세스들에 있는 단계들의 특정한 순서 또는 계층 구조가 재배열될 수 있다는 것을 이해하도록 한다. 첨부된 방법 청구항들은 샘플 순서로 다양한 단계들의 엘리먼트들을 제공하지만 제시된 특정한 순서 또는 계층 구조에 한정되는 것을 의미하지는 않는다.

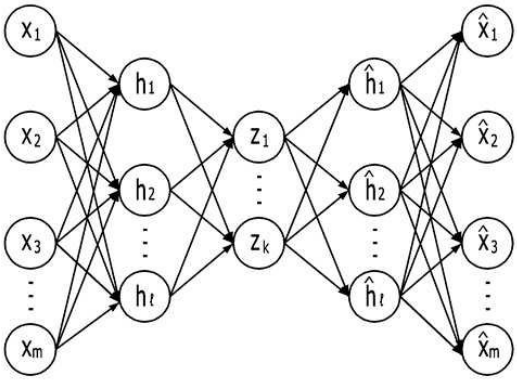
[0215] 제시된 실시예들에 대한 설명은 임의의 본 개시의 기술 분야에서 통상의 지식을 가진 자가 본 개시를 이용하거나 또는 실시할 수 있도록 제공된다. 이러한 실시예들에 대한 다양한 변형들은 본 개시의 기술 분야에서 통상의 지식을 가진 자에게 명백할 것이며, 여기에 정의된 일반적인 원리들은 본 개시의 범위를 벗어남이 없이 다른 실시예들에 적용될 수 있다. 그리하여, 본 개시는 여기에 제시된 실시예들로 한정되는 것이 아니라, 여기에 제시된 원리들 및 신규한 특징들과 일관되는 최광의의 범위에서 해석되어야 할 것이다.

도면

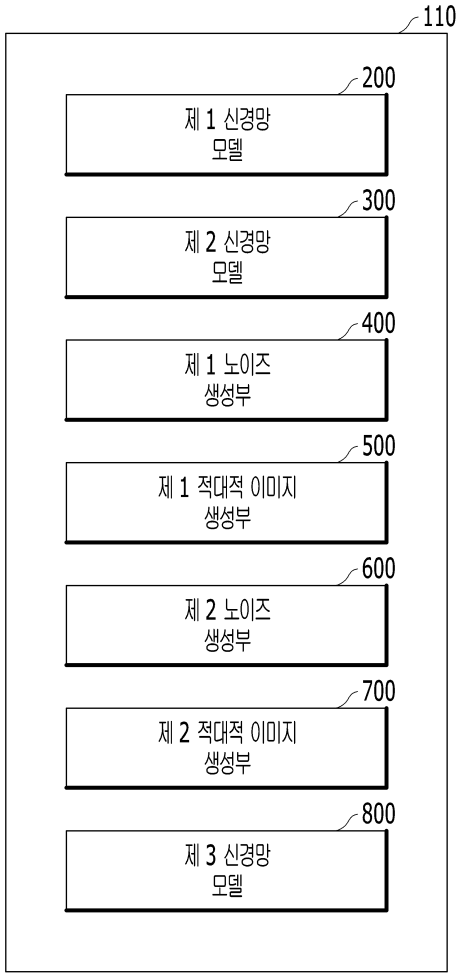
도면1


100
프로세서
110
메모리
130
네트워크부
150

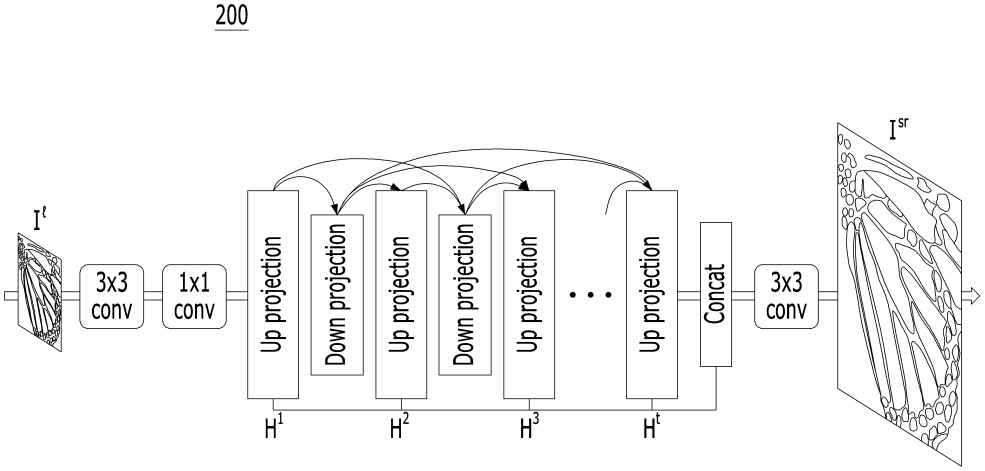
도면2


x1
x2
x3
xm
h1
h2
ht
z1
zk
ĥ1
ĥ2
ĥt
x̂1
x̂2
x̂3
x̂m

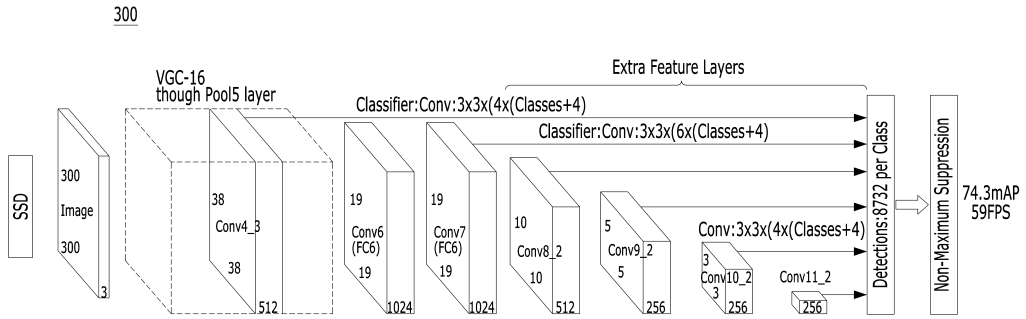
도면3


110
200
제 1 신경망
모델
300
제 2 신경망
모델
400
제 1 노이즈
생성부
500
제 1 적대적 이미지
생성부
600
제 2 노이즈
생성부
700
제 2 적대적 이미지
생성부
800
제 3 신경망
모델

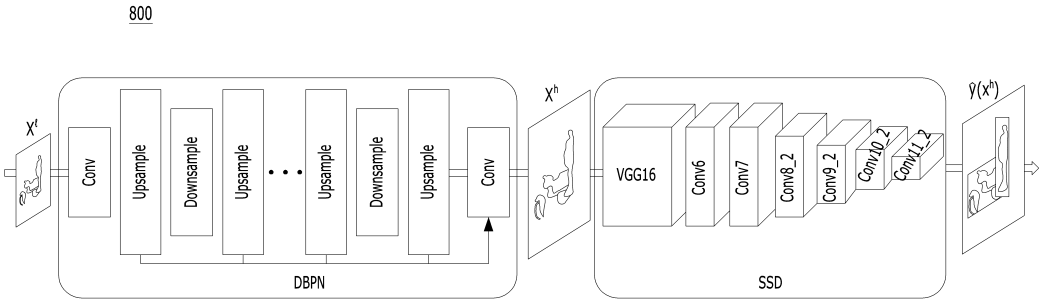
도면4


200
I^t
3x3 conv
1x1 conv
Up projection
Down projection
Up projection
Down projection
Up projection
Up projection
Concat
3x3 conv
I^{sr}
H^1
H^2
H^3
H^t

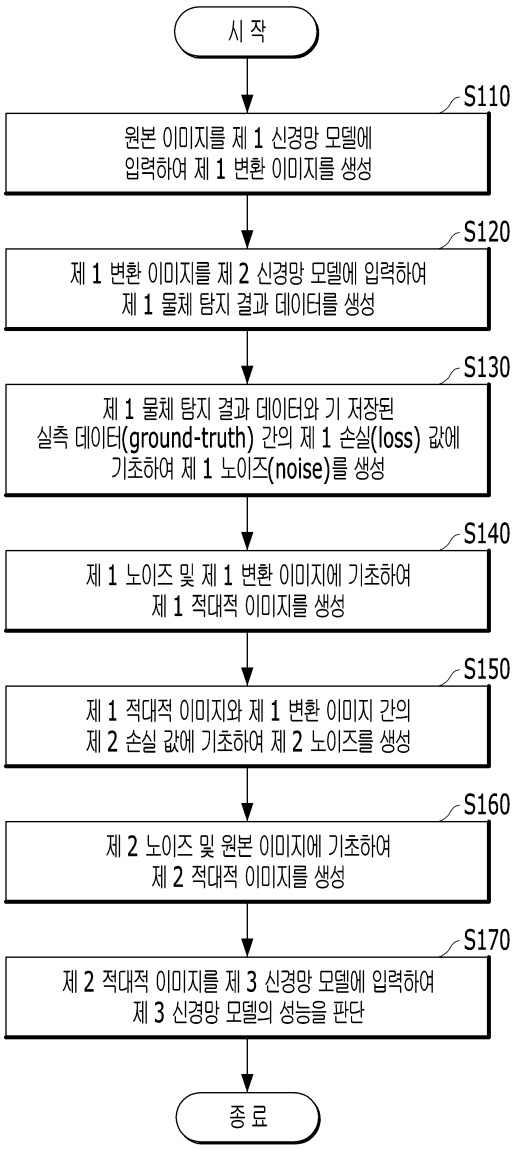
도면5


300
SSD
300
Image
300
3
VGC-16
though Pool5 layer
38
Conv4_3
38
512
19
Conv6
(FC6)
19
1024
19
Conv7
(FC6)
19
1024
10
Conv8_2
10
512
5
Conv9_2
5
256
3
Conv10_2
3
256
Conv11_2
256
Extra Feature Layers
Classifier:Conv:3x3x(4x(Classes+4)
Classifier:Conv:3x3x(6x(Classes+4)
Conv:3x3x(4x(Classes+4)
Detections:8732 per Class
Non-Maximum Suppression
74.3mAP
59FPS

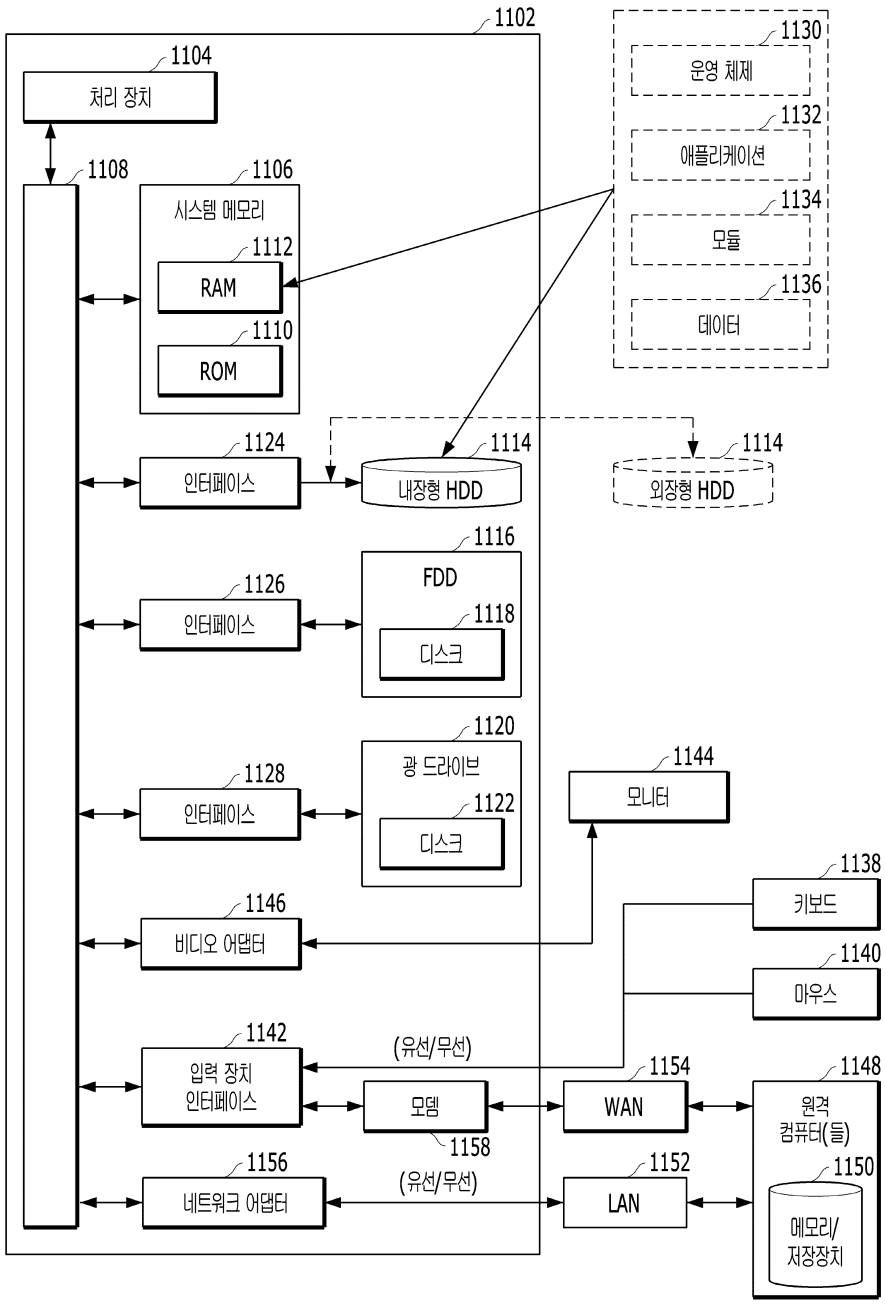
도면6


800
X^t
Conv
Upsample
Downsample
Upsample
Upsample
Downsample
Upsample
Conv
DBPN
X^h
VGG16
Conv6
Conv7
Conv8_2
Conv9_2
Conv10_2
Conv11_2
SSD
$\hat{Y}(X^h)$

도면7


시작
S110
원본 이미지를 제 1 신경망 모델에
입력하여 제 1 변환 이미지를 생성
S120
제 1 변환 이미지를 제 2 신경망 모델에 입력하여
제 1 물체 탐지 결과 데이터를 생성
S130
제 1 물체 탐지 결과 데이터와 기 저장된
실측 데이터(ground-truth) 간의 제 1 손실(loss) 값에
기초하여 제 1 노이즈(noise)를 생성
S140
제 1 노이즈 및 제 1 변환 이미지에 기초하여
제 1 적대적 이미지를 생성
S150
제 1 적대적 이미지와 제 1 변환 이미지 간의
제 2 손실 값에 기초하여 제 2 노이즈를 생성
S160
제 2 노이즈 및 원본 이미지에 기초하여
제 2 적대적 이미지를 생성
S170
제 2 적대적 이미지를 제 3 신경망 모델에 입력하여
제 3 신경망 모델의 성능을 판단
종료

도면8


1102
1104
처리 장치
1108
1106
시스템 메모리
1112
RAM
1110
ROM
1130
운영 체제
1132
애플리케이션
1134
모듈
1136
데이터
1124
인터페이스
1114
내장형 HDD
1114
외장형 HDD
1116
FDD
1126
인터페이스
1118
디스크
1120
광 드라이브
1128
인터페이스
1122
디스크
1144
모니터
1138
키보드
1146
비디오 어댑터
1140
마우스
1142
입력 장치 인터페이스
(유선/무선)
1154
WAN
1148
원격 컴퓨터(들)
모뎀
1158
1156
네트워크 어댑터
(유선/무선)
1152
LAN
1150
메모리/ 저장장치