



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0147895
(43) 공개일자 2023년10월24일

(51) 국제특허분류(Int. Cl.)
G06N 3/08 (2023.01) G06N 3/04 (2023.01)
(52) CPC특허분류
G06N 3/088 (2023.01)
G06N 3/047 (2023.01)
(21) 출원번호 10-2022-0046814
(22) 출원일자 2022년04월15일
심사청구일자 2022년04월15일

(71) 출원인
세종대학교산학협력단
서울특별시 광진구 능동로 209 (군자동, 세종대학교)
(72) 발명자
윤주범
서울특별시 송파구 충민로4길 19, 704동 401호(장지동, 송파파인타운7단지)
문현준
서울특별시 강북구 삼양로42길 15(미아동)
(74) 대리인
두호특허법인

전체 청구항 수 : 총 22 항

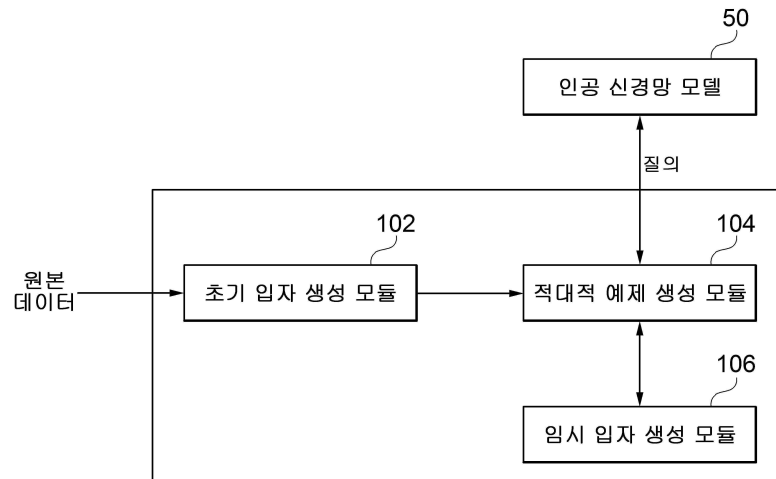
(54) 발명의 명칭 인공 신경망 모델의 적대적 공격 장치 및 방법

(57) 요약

인공 신경망 모델의 적대적 공격 장치 및 방법이 개시된다. 개시되는 일 실시예에 따른 적대적 공격 장치는, 인공 신경망 모델에 대한 적대적 공격을 수행하기 위한 적대적 공격 장치로서, 원본 데이터에 노이즈를 추가하여 초기 입자들을 생성하는 초기 입자 생성 모듈 및 초기 입자들을 인공 신경망 모델로 입력하고, 인공 신경망 모델의 출력값에 기반하여 각 초기 입자에 대한 적합도를 산출하며, 산출한 적합도에 기반하여 적대적 예제를 생성하는 적대적 예제 생성 모듈을 포함한다.

대표도 - 도1

100



(72) 발명자
서성관
 서울특별시 광진구 면목로5길 30-3, 301호(군자동)

손배훈
 서울특별시 광진구 광나루로13길 13, 304호(군자동)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711120086
과제번호	2020-0-01602
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	대학ICT연구센터지원사업
연구과제명	지능형 사이버 위협 대응 기술 개발 및 인력양성
기 여 율	1/1
과제수행기관명	송실대학교 산학협력단
연구기간	2020.07.01 ~ 2024.12.31

명세서

청구범위

청구항 1

인공 신경망 모델에 대한 적대적 공격을 수행하기 위한 적대적 공격 장치로서,

원본 데이터에 노이즈를 추가하여 초기 입자들을 생성하는 초기 입자 생성 모듈; 및

상기 초기 입자들을 상기 인공 신경망 모델로 입력하고, 상기 인공 신경망 모델의 출력값에 기반하여 각 초기 입자에 대한 적합도를 산출하며, 상기 산출한 적합도에 기반하여 적대적 예제를 생성하는 적대적 예제 생성 모듈을 포함하는, 적대적 공격 장치.

청구항 2

청구항 1에 있어서,

상기 적대적 예제 생성 모듈은,

하기 수학적식에 의해 초기 입자에 대한 적합도($Fitness_{score}$)를 산출하는, 적대적 공격 장치.

(수학적식)

$$Fitness_{score} = \{Z(x')_T - \operatorname{argmax}(Z(x')_{i \neq T})\}$$

x' : 초기 입자(노이즈가 추가된 원본 데이터)

$Z(x')$: 초기 입자의 입력에 대한 인공 신경망 모델의 출력 값

$Z(x')_T$: 초기 입자를 입력하였을 때 인공 신경망 모델의 출력 값이 적대적 예제를 통해 목표한 분류(T)로 나올 확률 값

$Z(x')_{i \neq T}$: 초기 입자를 입력하였을 때 인공 신경망 모델의 출력 값이 적대적 예제를 통해 목표한 분류 이외의 분류(i)로 나올 확률 값

$\operatorname{argmax}()$: ()를 최대화하는 인자를 찾는 함수

청구항 3

청구항 1에 있어서,

상기 적대적 예제 생성 모듈은,

상기 초기 입자를 기 설정된 탐색 공간 상에서 반복적으로 이동시키고, 이동된 각 위치에서 해당 초기 입자의 적합도를 각각 산출하는, 적대적 공격 장치.

청구항 4

청구항 3에 있어서,

상기 적대적 예제 생성 모듈은,

상기 탐색 공간 내에서 상기 초기 입자를 반복 이동시킬 때, 다음 번 이동에 따른 초기 입자의 위치(x_i^{t+1}) 및

속도(v_i^{t+1})를 하기의 수학식을 통해 산출하는, 적대적 공격 장치.

(수학식)

$$x_i^{t+1} = x_i^t + v_i^{t+1}$$

$$v_i^{t+1} = wv_i^t + c_1r_1(P_i - x_i^t) + c_2r_2(P_g - x_i^t)$$

x_i : i번째 초기 입자의 위치

t : 반복 횟수

v_i : i번째 초기 입자의 속도

w : 관성 가중치

c_1, c_2 : 기 설정된 상수

r_1, r_2 : 기 설정된 균일 분포 난수

P_i : i번째 초기 입자가 경험한 최상의 위치

P_g : 초기 입자들의 군집 전체의 최상의 위치

청구항 5

청구항 4에 있어서,

상기 적대적 예제 생성 모듈은,

상기 초기 입자들을 반복적으로 이동시킬 때, 상기 관성 가중치(w)를 선형적으로 감소시키면서 반복 이동시키는, 적대적 공격 장치.

청구항 6

청구항 5에 있어서,

상기 적대적 예제 생성 모듈은,

하기 수학식을 통해 상기 관성 가중치(w)를 선형적으로 감소시키는, 적대적 공격 장치.

(수학식)

$$w = w_{max} - t \cdot \frac{w_{max} - w_{min}}{t_{max}}$$

t : 반복 횟수

t_{max} : 기 설정된 반복 최대 횟수

w_{max} : 기 설정된 최대 관성 가중치

w_{min} : 기 설정된 최소 관성 가중치

청구항 7

청구항 4에 있어서,

상기 적대적 예제 생성 모듈은,

상기 초기 입자를 반복 이동시킬 때, 다음 번 이동에 따른 초기 입자의 위치(x_i^{t+1})를 하기 수학식에 의해 제한하여 이동시키는, 적대적 공격 장치.

(수학식)

$$x_i^{t+1} = \text{clip}(x_i^t + v_i^{t+1}, x - B, x + B)$$

B : 기 설정된 최대 변경 상한

Clip(리스트, 최소값, 최대값): 리스트에 있는 값을 최소값과 최대값 사이의 값으로 제한하는 함수

청구항 8

청구항 3에 있어서,

상기 적대적 예제 생성 모듈은,

상기 초기 입자의 이동을 반복적으로 수행하면서 상기 초기 입자의 적합도를 확인하여 기 설정된 임시 입자 생성 이벤트가 발생하는지 여부를 확인하는, 적대적 공격 장치.

청구항 9

청구항 8에 있어서,

상기 적대적 공격 장치는,

상기 임시 입자 생성 이벤트가 발생하는 경우, 상기 초기 입자들에 기반하여 임시 입자를 추가로 생성하는 임시 입자 생성 모듈을 더 포함하고,

상기 적대적 예제 생성 모듈은, 상기 생성된 임시 입자에 대해 적합도를 산출하는, 적대적 공격 장치.

청구항 10

청구항 9에 있어서,

상기 임시 입자 생성 모듈은,

상기 초기 입자들 중 제1 초기 입자 및 제2 초기 입자를 랜덤하게 선택하고, 선택된 상기 제1 초기 입자 및 상기 제2 초기 입자의 임의의 분할 지점을 기준으로 상기 제1 초기 입자의 일부 및 상기 제2 초기 입자의 일부를 붙여서 임시 입자를 생성하는, 적대적 공격 장치.

청구항 11

청구항 9에 있어서,

상기 적대적 예제 생성 모듈은,

상기 산출한 적합도가 기 설정된 임계 적합도 이상이 되는 초기 입자 또는 임시 입자를 상기 인공 신경망 모델에 대해 적대적 공격을 수행하기 위한 적대적 예제로 추출하는, 적대적 공격 장치.

청구항 12

하나 이상의 프로세서들, 및

상기 하나 이상의 프로세서들에 의해 실행되는 하나 이상의 프로그램들을 저장하는 메모리를 구비한 컴퓨팅 장치에서 수행되고, 인공 신경망 모델에 대한 적대적 공격을 수행하기 위한 적대적 공격 방법으로서,

원본 데이터에 노이즈를 추가하여 초기 입자들을 생성하는 단계; 및

상기 초기 입자들을 상기 인공 신경망 모델로 입력하고, 상기 인공 신경망 모델의 출력값에 기반하여 각 초기 입자에 대한 적합도를 산출하며, 상기 산출한 적합도에 기반하여 적대적 예제를 생성하는 단계를 포함하는, 적대적 공격 방법.

청구항 13

청구항 12에 있어서,

하기 수학적식에 의해 초기 입자에 대한 적합도(Fitness_{score})를 산출하는, 적대적 공격 장치.

(수학적식)

$$\text{Fitness}_{\text{score}} = \{Z(x')_T - \text{argmax}(Z(x')_{i \neq T})\}$$

x' : 초기 입자(노이즈가 추가된 원본 데이터)

$Z(x')$: 초기 입자의 입력에 대한 인공 신경망 모델의 출력 값

$Z(x')_T$: 초기 입자를 입력하였을 때 인공 신경망 모델의 출력 값이 적대적 예제를 통해 목표한 분류(T)로 나올 확률 값

$Z(x')_{i \neq T}$: 초기 입자를 입력하였을 때 인공 신경망 모델의 출력 값이 적대적 예제를 통해 목표한 분류 이외의 분류(i)로 나올 확률 값

$\text{argmax}()$: ()를 최대화하는 인자를 찾는 함수

청구항 14

청구항 12에 있어서,

상기 적대적 예제를 생성하는 단계는,

상기 초기 입자를 기 설정된 탐색 공간 상에서 반복적으로 이동시키고, 이동된 각 위치에서 해당 초기 입자의 적합도를 각각 산출하는 단계를 포함하는, 적대적 공격 방법.

청구항 15

청구항 14에 있어서,

상기 반복적으로 이동시키는 단계는,

상기 탐색 공간 내에서 상기 초기 입자를 반복 이동시킬 때, 다음 번 이동에 따른 초기 입자의 위치(x_i^{t+1}) 및 속도(v_i^{t+1})를 하기의 수학적식을 통해 산출하는, 적대적 공격 방법.

(수학적식)

$$x_i^{t+1} = x_i^t + v_i^{t+1}$$

$$v_i^{t+1} = wv_i^t + c_1r_1(P_i - x_i^t) + c_2r_2(P_g - x_i^t)$$

x_i : i번째 초기 입자의 위치

t : 반복 횟수

v_i : i 번째 초기 입자의 속도

w : 관성 가중치

c_1, c_2 : 기 설정된 상수

r_1, r_2 : 기 설정된 균일 분포 난수

P_i : i 번째 초기 입자가 경험한 최상의 위치

P_g : 초기 입자들의 군집 전체의 최상의 위치

청구항 16

청구항 15에 있어서,

상기 초기 입자들을 반복적으로 이동시킬 때, 상기 관성 가중치(w)를 선형적으로 감소시키면서 반복 이동시키는, 적대적 공격 방법.

청구항 17

청구항 16에 있어서,

하기 수학적식을 통해 상기 관성 가중치(w)를 선형적으로 감소시키는, 적대적 공격 방법.

(수학적식)

$$W = W_{max} - t \cdot \frac{W_{max} - W_{min}}{t_{max}}$$

t : 반복 횟수

t_{max} : 기 설정된 반복 최대 횟수

W_{max} : 기 설정된 최대 관성 가중치

W_{min} : 기 설정된 최소 관성 가중치

청구항 18

청구항 15에 있어서,

상기 초기 입자를 반복 이동시킬 때, 다음 번 이동에 따른 초기 입자의 위치(x_i^{t+1})를 하기 수학적식에 의해 제한하여 이동시키는, 적대적 공격 방법.

(수학적식)

$$x_i^{t+1} = clip(x_i^t + v_i^{t+1}, x - B, x + B)$$

B : 기 설정된 최대 변경 상한

Clip(리스트, 최소값, 최대값): 리스트에 있는 값을 최소값과 최대값 사이의 값으로 제한하는 함수

청구항 19

청구항 14에 있어서,

상기 적대적 예제를 생성하는 단계는,

상기 초기 입자의 이동을 반복적으로 수행하면서 상기 초기 입자의 적합도를 확인하여 기 설정된 임시 입자 생성 이벤트가 발생하는지 여부를 확인하는 단계;

상기 임시 입자 생성 이벤트가 발생하는 경우, 상기 초기 입자들에 기반하여 임시 입자를 추가로 생성하는 단계; 및

상기 생성된 임시 입자에 대해 적합도를 산출하는 단계를 더 포함하는, 적대적 공격 장치.

청구항 20

청구항 19에 있어서,

상기 임시 입자를 추가로 생성하는 단계는,

상기 초기 입자들 중 제1 초기 입자 및 제2 초기 입자를 랜덤하게 선택하고, 선택된 상기 제1 초기 입자 및 상기 제2 초기 입자의 임의의 분할 지점을 기준으로 상기 제1 초기 입자의 일부 및 상기 제2 초기 입자의 일부를 붙여서 임시 입자를 생성하는, 적대적 공격 방법.

청구항 21

청구항 19에 있어서,

상기 적대적 예제를 생성하는 단계는,

상기 산출한 적합도가 기 설정된 임계 적합도 이상이 되는 초기 입자 또는 임시 입자를 상기 인공 신경망 모델에 대해 적대적 공격을 수행하기 위한 적대적 예제로 추출하는 단계를 더 포함하는, 적대적 공격 방법.

청구항 22

비일시적 컴퓨터 판독 가능한 저장 매체(non-transitory computer readable storage medium)에 저장된 컴퓨터 프로그램으로서,

상기 컴퓨터 프로그램은 하나 이상의 명령어들을 포함하고, 인공 신경망 모델에 대한 적대적 공격을 수행하기 위한 것이며, 상기 명령어들은 하나 이상의 프로세서들을 갖는 컴퓨팅 장치에 의해 실행될 때, 상기 컴퓨팅 장치로 하여금,

원본 데이터에 노이즈를 추가하여 초기 입자들을 생성하는 단계; 및

상기 초기 입자들을 상기 인공 신경망 모델로 입력하고, 상기 인공 신경망 모델의 출력값에 기반하여 각 초기 입자에 대한 적합도를 산출하며, 상기 산출한 적합도에 기반하여 적대적 예제를 생성하는 단계를 수행하도록 하는, 컴퓨터 프로그램.

발명의 설명**기술 분야**

본 발명의 실시예는 인공 신경망 모델에 대한 적대적 공격 기술과 관련된다.

[0001]

배경 기술

- [0003] 적대적 공격(Adversarial Attack)이란 데이터에 사람이 인식할 수 없는 노이즈를 추가하여 인간의 인식에는 데이터에 아무런 변화가 없지만, 분류에 사용되는 딥 러닝 모델은 이를 잘못 분류하도록 하는 공격을 의미한다. 이때, 원 본 데이터를 적대적 공격하여 생성된 데이터를 적대적 예제(Adversarial Example)라고 하며, 딥 러닝 모델이 적대적 예제를 오분류 하였을 때 공격이 성공했다고 한다.
- [0004] 적대적 공격은 크게 딥 러닝 모델의 매개 변수, 손실 함수, 구조 등 그 내부 정보를 사용하는 화이트 박스 공격(white box attack)과 딥 러닝 모델의 내부 정보 없이 오직 딥 러닝 모델의 입력과 출력 값만을 가지고 수행하는 블랙 박스 공격(black box attack)으로 구분할 수 있다. 그 중에서도 딥 러닝 모델의 내부 정보를 알 수 없는 블랙박스 적대적 공격은 더욱 실용적이고 어려운 공격으로 평가된다.
- [0005] 기존의 블랙박스 적대적 공격은 주로 모집단 기반 최적화 알고리즘에 기반하나, 이러한 모집단 기반 알고리즘은 크게 두 가지 문제가 있다. 첫째, 적대적 공격에 필요한 연산 자원이 많이 필요하다는 문제가 있다. 즉, 딥 러닝 모델에 대한 입력 데이터의 질의와 그에 대한 출력 값을 기반으로 최적화를 수행하는 특성으로 인해, 딥 러닝 모델에 최소 수천 회에서 수만 회까지 질의가 필요하다. 이는 곧 막대한 연산 자원을 필요로 할 뿐만 아니라, 비정상적인 질의 횟수로 인해 공격 탐지 위험성 또한 매우 높게 된다.
- [0006] 두 번째로, 기존의 모집단 기반 알고리즘은 공격 성공률이 낮다는 문제가 있다. 즉, 부족한 정보로 인해 블랙박스 적대적 공격의 성공률은 소모 자원에 비해 낮다. 한편, 신뢰도 높은 적대적 공격을 위해서는 높은 공격 성공률이 필수적이나, 이는 적은 자원으로 달성하기 어렵다는 문제가 있다. 따라서, 효율적이고 신뢰도 높은 블랙박스 적대적 공격을 위해서는 딥 러닝 모델에 대해 적은 질의로도 높은 공격 성공률을 가지는 공격 방법이 요구된다.
- [0007] 딥 러닝 모델은 의료, 자율 주행, 얼굴 인식 등 오 작동 시 매우 위험하거나 보안 상의 치명적인 문제를 일으킬 수 있는 중요한 분야에 적용되기 때문에, 높은 강건성을 가져야 하며 이에 대해 적절히 평가 받아야 한다. 여기서, 적대적 공격은 딥 러닝 모델의 강건성을 평가하고 개선하는데 매우 중요한 역할을 하게 된다. 즉, 효율적인 적대적 예제 생성은 곧 효율적인 딥 러닝 모델의 강건성 평가로 이어지는 바, 높은 공격 성공률과 상대적으로 적은 자원으로도 공격을 수행할 수 있는 적대적 공격 방법이 필요하다.

선행기술문헌

특허문헌

- [0009] (특허문헌 0001) 한국공개특허공보 제10-2022-0025455호(2022.03.03)

발명의 내용

해결하려는 과제

- [0010] 본 발명의 실시예는 상대적으로 적은 자원으로 공격 성공률을 높일 수 있는 적대적 공격 방법 및 장치를 제공하기 위한 것이다.

과제의 해결 수단

- [0012] 개시되는 일 실시예에 따른 적대적 공격 장치는, 인공 신경망 모델에 대한 적대적 공격을 수행하기 위한 적대적 공격 장치로서, 원본 데이터에 노이즈를 추가하여 초기 입자들을 생성하는 초기 입자 생성 모듈; 및 상기 초기 입자들을 상기 인공 신경망 모델로 입력하고, 상기 인공 신경망 모델의 출력값에 기반하여 각 초기 입자에 대한 적합도를 산출하며, 상기 산출한 적합도에 기반하여 적대적 예제를 생성하는 적대적 예제 생성 모듈을 포함한다.
- [0013] 상기 적대적 예제 생성 모듈은, 하기 수학식에 의해 초기 입자에 대한 적합도(Fitness_{score})를 산출할 수 있다.

- [0014] (수학식)
- [0015]
$$\text{Fitness}_{\text{score}} = \{Z(x')_T - \text{argmax}(Z(x')_{i \neq T})\}$$
- [0016] x' : 초기 입자(노이즈가 추가된 원본 데이터)
- [0017] $Z(x')$: 초기 입자의 입력에 대한 인공 신경망 모델의 출력 값
- [0018] $Z(x')_T$: 초기 입자를 입력하였을 때 인공 신경망 모델의 출력 값이 적대적 예제를 통해 목표한 분류(T)로 나올 확률 값
- [0019] $Z(x')_{i \neq T}$: 초기 입자를 입력하였을 때 인공 신경망 모델의 출력 값이 적대적 예제를 통해 목표한 분류 이외의 분류(i)로 나올 확률 값
- [0020] $\text{argmax}()$: ()를 최대화하는 인자를 찾는 함수
- [0021] 상기 적대적 예제 생성 모듈은, 상기 초기 입자를 기 설정된 탐색 공간 상에서 반복적으로 이동시키고, 이동된 각 위치에서 해당 초기 입자의 적합도를 각각 산출할 수 있다.
- [0022] 상기 적대적 예제 생성 모듈은, 상기 탐색 공간 내에서 상기 초기 입자를 반복 이동시킬 때, 다음 번 이동에 따른 초기 입자의 위치(x_i^{t+1}) 및 속도(v_i^{t+1})를 하기의 수학식을 통해 산출할 수 있다.
- [0023] (수학식)
- [0024]
$$x_i^{t+1} = x_i^t + v_i^{t+1}$$
- [0025]
$$v_i^{t+1} = wv_i^t + c_1r_1(P_i - x_i^t) + c_2r_2(P_g - x_i^t)$$
- [0026] x_i : i번째 초기 입자의 위치
- [0027] t : 반복 횟수
- [0028] v_i : i번째 초기 입자의 속도
- [0029] w : 관성 가중치
- [0030] c_1, c_2 : 기 설정된 상수
- [0031] r_1, r_2 : 기 설정된 균일 분포 난수
- [0032] P_i : i번째 초기 입자가 경험한 최상의 위치
- [0033] P_g : 초기 입자들의 군집 전체의 최상의 위치
- [0034] 상기 적대적 예제 생성 모듈은, 상기 초기 입자들을 반복적으로 이동시킬 때, 상기 관성 가중치(w)를 선형적으로 감소시키면서 반복 이동시킬 수 있다.
- [0035] 상기 적대적 예제 생성 모듈은, 하기 수학식을 통해 상기 관성 가중치(w)를 선형적으로 감소시킬 수 있다.
- [0036] (수학식)
- [0037]
$$w = w_{\max} - t \cdot \frac{w_{\max} - w_{\min}}{t_{\max}}$$
- [0038] t : 반복 횟수
- [0039] $t_{\max}$: 기 설정된 반복 최대 횟수
- [0040] $w_{\max}$: 기 설정된 최대 관성 가중치
- [0041] $w_{\min}$: 기 설정된 최소 관성 가중치

- [0042] 상기 적대적 예제 생성 모듈은, 상기 초기 입자를 반복 이동시킬 때, 다음 번 이동에 따른 초기 입자의 위치(x_i^{t+1})를 하기 수학식에 의해 제한하여 이동시킬 수 있다.
- [0043] (수학식)
- [0044]
$$x_i^{t+1} = \text{clip}(x_i^t + v_i^{t+1}, x - B, x + B)$$
- [0045] B : 기 설정된 최대 변경 상한
- [0046] Clip(리스트, 최소값, 최대값): 리스트에 있는 값을 최소값과 최대값 사이의 값으로 제한하는 함수
- [0047] 상기 적대적 예제 생성 모듈은, 상기 초기 입자의 이동을 반복적으로 수행하면서 상기 초기 입자의 적합도를 확인하여 기 설정된 임시 입자 생성 이벤트가 발생하는지 여부를 확인할 수 있다.
- [0048] 상기 적대적 공격 장치는, 상기 임시 입자 생성 이벤트가 발생하는 경우, 상기 초기 입자들에 기반하여 임시 입자를 추가로 생성하는 임시 입자 생성 모듈을 더 포함하고, 상기 적대적 예제 생성 모듈은, 상기 생성된 임시 입자에 대해 적합도를 산출할 수 있다.
- [0049] 상기 임시 입자 생성 모듈은, 상기 초기 입자들 중 제1 초기 입자 및 제2 초기 입자를 랜덤하게 선택하고, 선택된 상기 제1 초기 입자 및 상기 제2 초기 입자의 임의의 분할 지점을 기준으로 상기 제1 초기 입자의 일부 및 상기 제2 초기 입자의 일부를 붙여서 임시 입자를 생성할 수 있다.
- [0050] 상기 적대적 예제 생성 모듈은, 상기 산출한 적합도가 기 설정된 임계 적합도 이상이 되는 초기 입자 또는 임시 입자를 상기 인공 신경망 모델에 대해 적대적 공격을 수행하기 위한 적대적 예제로 추출할 수 있다.
- [0051] 개시되는 일 실시예에 따른 적대적 공격 방법은, 하나 이상의 프로세서들, 및 상기 하나 이상의 프로세서들에 의해 실행되는 하나 이상의 프로그램들을 저장하는 메모리를 구비한 컴퓨팅 장치에서 수행되고, 인공 신경망 모델에 대한 적대적 공격을 수행하기 위한 적대적 공격 방법으로서, 원본 데이터에 노이즈를 추가하여 초기 입자들을 생성하는 단계; 및 상기 초기 입자들을 상기 인공 신경망 모델로 입력하고, 상기 인공 신경망 모델의 출력값에 기반하여 각 초기 입자에 대한 적합도를 산출하며, 상기 산출한 적합도에 기반하여 적대적 예제를 생성하는 단계를 포함한다.
- [0052] 하기 수학식에 의해 초기 입자에 대한 적합도(Fitness_{score})를 산출할 수 있다.
- [0053] (수학식)
- [0054]
$$\text{Fitness}_{\text{score}} = \{Z(x')_T - \text{argmax}(Z(x')_{i \neq T})\}$$
- [0055] x' : 초기 입자(노이즈가 추가된 원본 데이터)
- [0056] $Z(x')$: 초기 입자의 입력에 대한 인공 신경망 모델의 출력 값
- [0057] $Z(x')_T$: 초기 입자를 입력하였을 때 인공 신경망 모델의 출력 값이 적대적 예제를 통해 목표한 분류(T)로 나올 확률 값
- [0058] $Z(x')_{i \neq T}$: 초기 입자를 입력하였을 때 인공 신경망 모델의 출력 값이 적대적 예제를 통해 목표한 분류 이외의 분류(i)로 나올 확률 값
- [0059] $\text{argmax}()$: ()를 최대화하는 인자를 찾는 함수
- [0060] 상기 적대적 예제를 생성하는 단계는, 상기 초기 입자를 기 설정된 탐색 공간 상에서 반복적으로 이동시키고, 이동된 각 위치에서 해당 초기 입자의 적합도를 각각 산출하는 단계를 포함할 수 있다.
- [0061] 상기 반복적으로 이동시키는 단계는, 상기 탐색 공간 내에서 상기 초기 입자를 반복 이동시킬 때, 다음 번 이동에 따른 초기 입자의 위치(x_i^{t+1}) 및 속도(v_i^{t+1})를 하기의 수학식을 통해 산출할 수 있다.
- [0062] (수학식)
- [0063]
$$x_i^{t+1} = x_i^t + v_i^{t+1}$$

$$v_i^{t+1} = wv_i^t + c_1r_1(P_i - x_i^t) + c_2r_2(P_g - x_i^t)$$

x_i : i번째 초기 입자의 위치

t : 반복 횟수

v_i : i번째 초기 입자의 속도

w : 관성 가중치

c_1, c_2 : 기 설정된 상수

r_1, r_2 : 기 설정된 균일 분포 난수

P_i : i번째 초기 입자가 경험한 최상의 위치

P_g : 초기 입자들의 군집 전체의 최상의 위치

상기 초기 입자들을 반복적으로 이동시킬 때, 상기 관성 가중치(w)를 선형적으로 감소시키면서 반복 이동시킬 수 있다.

하기 수학적식을 통해 상기 관성 가중치(w)를 선형적으로 감소시킬 수 있다.

(수학적식)

$$w = w_{max} - t \cdot \frac{w_{max} - w_{min}}{t_{max}}$$

t : 반복 횟수

t_{max} : 기 설정된 반복 최대 횟수

w_{max} : 기 설정된 최대 관성 가중치

w_{min} : 기 설정된 최소 관성 가중치

상기 초기 입자를 반복 이동시킬 때, 다음 번 이동에 따른 초기 입자의 위치(x_i^{t+1})를 하기 수학적식에 의해 제한하여 이동시킬 수 있다.

(수학적식)

$$x_i^{t+1} = clip(x_i^t + v_i^{t+1}, x - B, x + B)$$

B : 기 설정된 최대 변경 상한

Clip(리스트, 최소값, 최대값): 리스트에 있는 값을 최소값과 최대값 사이의 값으로 제한하는 함수

상기 적대적 예제를 생성하는 단계는, 상기 초기 입자의 이동을 반복적으로 수행하면서 상기 초기 입자의 적합도를 확인하여 기 설정된 임시 입자 생성 이벤트가 발생하는지 여부를 확인하는 단계; 상기 임시 입자 생성 이벤트가 발생하는 경우, 상기 초기 입자들에 기반하여 임시 입자를 추가로 생성하는 단계; 및 상기 생성된 임시 입자에 대해 적합도를 산출하는 단계를 더 포함할 수 있다.

상기 임시 입자를 추가로 생성하는 단계는, 상기 초기 입자들 중 제1 초기 입자 및 제2 초기 입자를 랜덤하게 선택하고, 선택된 상기 제1 초기 입자 및 상기 제2 초기 입자의 임의의 분할 지점을 기준으로 상기 제1 초기 입자의 일부 및 상기 제2 초기 입자의 일부를 붙여서 임시 입자를 생성할 수 있다.

상기 적대적 예제를 생성하는 단계는, 상기 산출한 적합도가 기 설정된 임계 적합도 이상이 되는 초기 입자 또는 임시 입자를 상기 인공 신경망 모델에 대해 적대적 공격을 수행하기 위한 적대적 예제로 추출하는 단계를 더 포함할 수 있다.

발명의 효과

- [0090] 개시되는 실시예에 의하면, 원본 데이터에 노이즈를 추가하여 초기 입자를 생성하고, 초기 입자를 탐색 공간 상에서 반복적으로 이동시키면서 적합도를 산출함으로써, 적대적 공격을 수행하기 위한 최적의 적대적 예제를 추출할 수 있게 된다. 그리고, 임시 입자 생성 이벤트가 발생하면, 임시 입자를 추가적으로 생성함으로써 공격 탐지 위험을 감소시킬 수 있고 효율적인 적대적 공격을 수행할 수 있게 된다.

도면의 간단한 설명

- [0092] 도 1은 본 발명의 일 실시예에 따른 적대적 공격 장치의 구성을 나타낸 블록도
- 도 2는 본 발명의 일 실시예에서 탐색 공간 내에서 초기 입자를 이동시키는 상태를 나타낸 도면
- 도 3은 본 발명의 일 실시예에서 초기 입자의 이동을 반복하면서 최적의 위치에 도달하는 과정을 개략적으로 나타낸 도면
- 도 4는 본 발명의 일 실시예에서 초기 입자에 기반하여 임시 입자를 생성하는 상태를 개략적으로 나타낸 도면
- 도 5는 본 발명의 일 실시예에 따른 적대적 공격 방법을 나타낸 흐름도
- 도 6은 본 발명의 일 실시예에 따른 적대적 공격 방법을 개략적으로 나타낸 도면
- 도 7은 예시적인 실시예들에서 사용되기에 적합한 컴퓨팅 장치를 포함하는 컴퓨팅 환경을 예시하여 설명하기 위한 블록도

발명을 실시하기 위한 구체적인 내용

- [0093] 이하, 도면을 참조하여 본 발명의 구체적인 실시형태를 설명하기로 한다. 이하의 상세한 설명은 본 명세서에서 기술된 방법, 장치 및/또는 시스템에 대한 포괄적인 이해를 돕기 위해 제공된다. 그러나 이는 예시에 불과하며 본 발명은 이에 제한되지 않는다.
- [0094] 본 발명의 실시예들을 설명함에 있어서, 본 발명과 관련된 공지기술에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략하기로 한다. 그리고, 후술되는 용어들은 본 발명에서의 기능을 고려하여 정의된 용어들로서 이는 사용자, 운용자의 의도 또는 관례 등에 따라 달라질 수 있다. 그러므로 그 정의는 본 명세서 전반에 걸친 내용을 토대로 내려져야 할 것이다. 상세한 설명에서 사용되는 용어는 단지 본 발명의 실시예들을 기술하기 위한 것이며, 결코 제한적이어서는 안 된다. 명확하게 달리 사용되지 않는 한, 단수 형태의 표현은 복수 형태의 의미를 포함한다. 본 설명에서, "포함" 또는 "구비"와 같은 표현은 어떤 특성들, 숫자들, 단계들, 동작들, 요소들, 이들의 일부 또는 조합을 가리키기 위한 것이며, 기술된 것 이외에 하나 또는 그 이상의 다른 특성, 숫자, 단계, 동작, 요소, 이들의 일부 또는 조합의 존재 또는 가능성을 배제하도록 해석되어서는 안 된다.
- [0095] 또한, 제1, 제2 등의 용어는 다양한 구성 요소들을 설명하는데 사용될 수 있지만, 상기 구성 요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성 요소를 다른 구성 요소로부터 구별하는 목적으로 사용될 수 있다. 예를 들어, 본 발명의 권리 범위를 벗어나지 않으면서 제1 구성 요소는 제2 구성 요소로 명명될 수 있고, 유사하게 제2 구성 요소도 제1 구성 요소로 명명될 수 있다.
- [0097] 도 1은 본 발명의 일 실시예에 따른 적대적 공격 장치의 구성을 나타낸 블록도이다.
- [0098] 도 1을 참조하면, 적대적 공격 장치(100)는 초기 입자 생성 모듈(102), 적대적 예제 생성 모듈(104), 및 임시 입자 생성 모듈(106)을 포함할 수 있다. 적대적 공격 장치(100)는 인공 신경망 모델(50)에 대한 적대적 공격을 수행하기 위한 장치이다. 예시적인 실시예에서, 적대적 공격 장치(100)는 인공 신경망 모델(50)에 대해 블랙박스 적대적 공격을 수행하기 위한 장치일 수 있다.
- [0099] 초기 입자 생성 모듈(102)은 입력되는 원본 데이터에 노이즈(예를 들어, 가우시안 노이즈)를 추가하여 초기 입자를 생성할 수 있다. 여기서, 원본 데이터는 원본 데이터 셋이 저장된 데이터베이스로부터 입력 받을 수 있으나 이에 한정되는 것은 아니다. 이하에서는, 설명의 편의상 원본 데이터가 오디오 데이터인 것을 일 예로 설명

하기로 하나, 원본 데이터가 이에 한정되는 것은 아니며 원본 데이터는 이미지 데이터 등 다양한 형태의 데이터를 포함할 수 있다.

[0100] 원본 데이터가 오디오 데이터인 경우, 초기 입자 생성 모듈(102)은 원본 데이터에 백색 잡음을 추가하여 초기 입자를 생성할 수 있다. 여기서, 초기 입자의 생성은 원본 데이터에 백색 잡음을 추가하여 데이터 분포 공간 상에서 원본 데이터의 위치를 백색 잡음이 추가된 만큼 변경시키는 것일 수 있다. 예를 들어, 오디오 데이터에 백색 잡음을 추가하면 오디오 데이터가 원래 가지고 있던 진폭 및 주파수 등이 달라지게 되는데, 백색 잡음이 추가됨으로써 진폭 및 주파수 등이 달라진 오디오 데이터를 초기 입자라고 지칭할 수 있다.

[0101] 예시적인 실시예에서, 초기 입자 생성은 입자 군집 최적화(Particle Swarm Optimization : PSO) 기술을 통해 이루어질 수 있다. 초기 입자 생성 모듈(102)은 원본 데이터에 백색 잡음을 랜덤하게 추가하여 기 설정된 개수의 초기 입자를 생성할 수 있다.

[0102] 적대적 예제 생성 모듈(104)은 생성된 초기 입자들에 기반하여 인공 신경망 모델(50)에 대해 적대적 공격을 수행하기 위한 적대적 예제를 생성할 수 있다. 구체적으로, 적대적 예제 생성 모듈(104)은 생성된 초기 입자를 인공 신경망 모델(50)로 입력하고, 인공 신경망 모델(50)의 출력값에 기반하여 해당 초기 입자에 대한 적합도를 산출할 수 있다.

[0103] 여기서, 적합도란 해당 초기 입자가 인공 신경망 모델(50)에 대해 적대적 공격을 수행하기 위한 적대적 예제가 되는데 적합한지의 정도를 나타내는 것일 수 있다. 적대적 예제 생성 모듈(104)에서 초기 입자를 인공 신경망 모델(50)로 입력하여 그에 대한 출력 값을 획득하는 과정을 질의 절차라 할 수 있다.

[0104] 적대적 예제 생성 모듈(104)은 다음의 수학적 식 1을 통해 초기 입자에 대한 적합도(Fitness_{score})를 산출할 수 있다. 여기서, 적합도가 높을수록 해당 초기 입자가 적대적 예제가 되는데 적합함을 의미하게 된다.

[0105] (수학적 식 1)

$$[0106] \text{Fitness}_{\text{score}} = \{Z(x')_T - \text{argmax}(Z(x')_{i \neq T})\}$$

[0107] x' : 초기 입자(노이즈가 추가된 원본 데이터)

[0108] $Z(x')$: 초기 입자의 입력에 대한 인공 신경망 모델의 출력 값

[0109] $Z(x')_T$: 초기 입자를 입력하였을 때 인공 신경망 모델의 출력 값이 적대적 예제를 통해 목표한 분류(T)로 나올 확률 값

[0110] $Z(x')_{i \neq T}$: 초기 입자를 입력하였을 때 인공 신경망 모델의 출력 값이 적대적 예제를 통해 목표한 분류 이외의 분류(i)로 나올 확률 값

[0111] $\text{argmax}()$: ()를 최대화하는 인자를 찾는 함수

[0112] 적대적 예제 생성 모듈(104)은 초기 입자의 처음 위치 및 초기 입자의 처음 위치에서 산출된 적합도를 각각 저장할 수 있다. 또한, 적대적 예제 생성 모듈(104)은 초기 입자를 데이터 분포 공간 상에서 이동시키면서 적합도를 산출할 수 있다.

[0113] 즉, 적대적 예제 생성 모듈(104)은 데이터 분포 공간 상에서 초기 입자의 이동을 통해 초기 입자의 위치를 최적화 할 수 있다. 적대적 예제 생성 모듈(104)은 초기 입자를 반복적으로 이동시켜 초기 입자의 위치를 최적화 할 수 있다. 여기서, 초기 입자의 위치를 최적화 한다고 하는 것은 초기 입자가 적대적 예제가 되기 위해 최적화된 위치를 찾는 것을 의미할 수 있다. 적대적 예제 생성 모듈(104)은 초기 입자를 반복적으로 이동시켜 가면서 이동된 위치에서의 적합도를 산출한 후, 초기 입자의 위치 및 적합도를 갱신하여 저장할 수 있다. 그리고, 데이터 분포 공간은 초기 입자의 최적화 위치를 찾기 위한 공간이라는 의미에서 탐색 공간으로 지칭할 수 있다.

[0114] 도 2는 본 발명의 일 실시예에서 탐색 공간 내에서 초기 입자를 이동시키는 상태를 나타낸 도면이다. 여기서, 초기 입자의 이동이 반복됨에 따라 초기 입자의 위치는 다음의 수학적 식 2로 나타낼 수 있다.

[0115] (수학적 식 2)

$$[0116] x_i^{t+1} = x_i^t + v_i^{t+1}$$

- [0117] x_i : i번째 초기 입자의 위치
- [0118] t : 반복 횟수
- [0119] v_i : i번째 초기 입자의 속도
- [0120] 여기서, 탐색 공간이 n차원인 경우, x_i 및 v_i 는 각각 n차원 벡터로 나타낼 수 있다. 그리고, v_i 는 i번째 초기 입자의 속도로서, i번째 초기 입자가 이동하는 방향을 나타낸다.
- [0121] 한편, 초기 입자의 다음 번(t+1) 이동에 따른 초기 입자의 속도(v_i^{t+1})는 다음의 수학적 식 3으로 나타낼 수 있다.
- [0122] (수학적 식 3)
- [0123]
$$v_i^{t+1} = wv_i^t + c_1r_1(P_i - x_i^t) + c_2r_2(P_g - x_i^t)$$
- [0124] w : 관성 가중치
- [0125] c_1, c_2 : 기 설정된 상수
- [0126] r_1, r_2 : 기 설정된 균일 분포 난수
- [0127] P_i : i번째 초기 입자가 경험한 최상의 위치
- [0128] P_g : 초기 입자들의 군집 전체의 최상의 위치
- [0129] 즉, P_i 는 i번째 초기 입자가 반복적인 이동을 하면서 경험한 최적의 위치에서 가장 가까운 최상의 위치를 나타내고, P_g 는 초기 입자들의 군집 전체를 대상으로 최적의 위치에서 가장 가까운 최상의 위치를 나타낸다.
- [0130] 수학적 식 3에서, 우변의 첫 번째 항(wv_i^t)은 다음 속도에 대한 현재 속도의 영향을 조절하는 관성을 나타내며, 관성 가중치(w)에 의해 가중된다. 우변의 두 번째 항($c_1r_1(P_i - x_i^t)$)은 기억(memory)과 관련된 것으로 각 입자가 경험한 최상의 위치를 나타내는 것이며 기억이 속도에 영향을 미치는 범위를 결정하기 위해 c_1r_1 으로 가중치가 부여될 수 있다. 우변의 세 번째 항($c_2r_2(P_g - x_i^t)$)은 정보(information)와 관련된 것으로 입자 군집 전체의 최상의 위치를 나타내는 것이며 정보가 속도에 영향을 미치는 범위를 결정하기 위해 c_2r_2 로 가중치가 부여될 수 있다.
- [0131] 적대적 예제 생성 모듈(104)은 탐색 공간에서 초기 입자들을 반복적으로 이동시킬 때 마다 각 초기 입자들의 적합도를 산출하고, 초기 입자들의 위치 및 그에 따른 적합도를 갱신할 수 있다.
- [0132] 예시적인 실시예에서, 적대적 예제 생성 모듈(104)은 탐색 공간에서 초기 입자들을 반복적으로 이동시킬 때, 관성 가중치(w)를 선형적으로 감소시킬 수 있다. 관성 가중치(w)를 선형적으로 감소시키면서 초기 입자들을 반복적으로 이동시키면, 초기 입자가 반복 초기에는 더 많이 움직이고 반복 횟수가 많아질수록 덜 움직이게 된다. 이 경우, 초기 입자의 위치 최적화가 진행됨에 따라 정확한 섭동(perturbation) 제어가 필요한 적대적 예제를 생성하는데 적합하게 된다. 관성 가중치(w)는 초기 입자의 이동이 반복될수록 다음의 수학적 식 4와 같이 선형적으로 감소될 수 있다.
- [0133] (수학적 식 4)
- [0134]
$$w = w_{max} - t \cdot \frac{w_{max} - w_{min}}{t_{max}}$$
- [0135] t : 반복 횟수
- [0136] t_{max} : 기 설정된 반복 최대 횟수
- [0137] w_{max} : 기 설정된 최대 관성 가중치

- [0138] $w_{\min}$: 기 설정된 최소 관성 가중치
- [0139] 한편, 적대적 예제 생성 모듈(104)은 초기 입자를 이동시킬 때 원본 데이터에 과도한 섭동이 추가되는 것을 방지하기 위해 기 설정된 최대 변경 상한 내에서 이동되도록 할 수 있다. 적대적 예제 생성 모듈(104)은 초기 입자를 다음 번 위치로 이동시킬 때, 다음의 수학적 식 5에 의해 초기 입자가 기 설정된 최대 변경 상한 내에서 이동되도록 할 수 있다.
- [0140] (수학적 식 5)
- [0141]
$$x_i^{t+1} = \text{clip}(x_i^t + v_i^{t+1}, x - B, x + B)$$
- [0142] B : 기 설정된 최대 변경 상한
- [0143] Clip(리스트, 최소값, 최대값): 리스트에 있는 값을 최소값과 최대값 사이의 값으로 제한하는 함수
- [0144] 적대적 예제 생성 모듈(104)은 기 설정된 최대 반복 횟수까지 초기 입자를 이동시킬 수 있다. 또한, 적대적 예제 생성 모듈(104)은 초기 입자의 적합도가 기 설정된 임계 적합도 이상이 될 때까지(즉, 적대적 예제가 되기 위한 최적의 위치에 도달하거나 근접하게 도달할 때까지) 초기 입자를 반복적으로 이동시킬 수 있다. 도 3은 본 발명의 일 실시예에서 초기 입자의 이동을 반복하면서 최적의 위치에 도달하는 과정을 개략적으로 나타낸 도면이다.
- [0145] 적대적 예제 생성 모듈(104)은 초기 입자의 이동을 반복적으로 수행하면서 초기 입자의 적합도를 확인하여 임시 입자 생성 이벤트가 발생하는지 여부를 확인할 수 있다. 예를 들어, 적대적 예제 생성 모듈(104)은 기 설정된 최대 반복 횟수까지 초기 입자를 이동시켰어도 적합도가 임계 적합도에 도달하지 못하거나 초기 입자의 이동을 일정 횟수 이상 반복적으로 수행하였는데도 적합도가 더 이상 개선되지 않는 경우 임시 입자 생성 이벤트가 발생한 것으로 판단할 수 있다.
- [0146] 임시 입자 생성 모듈(106)은 임시 입자 생성 이벤트가 발생한 경우, 기 생성된 초기 입자에 기반하여 임시 입자를 생성할 수 있다.
- [0147] 구체적으로, 임시 입자 생성 모듈(106)은 초기 입자들 중에서 2개의 초기 입자를 랜덤하게 선택할 수 있다. 임시 입자 생성 모듈(106)은 선택된 2개의 초기 입자에 기반하여 임시 입자를 생성할 수 있다. 예를 들어, 도 4에 도시된 바와 같이, 초기 입자(P1) 및 초기 입자(P2)의 데이터의 임의의 분할 지점을 기준으로 초기 입자(P1)의 일부와 초기 입자(P2)의 일부를 붙여서 새로운 입자인 임시 입자(P3)를 생성할 수 있다. 여기서, 초기 입자(P1) 및 초기 입자(P2)는 일종의 부모 입자가 되고, 임시 입자(P3)는 자식 입자가 된다. 또한, 임시 입자 생성 모듈(106)은 생성된 임시 입자에 노이즈를 랜덤하게 추가할 수 있다.
- [0148] 임시 입자 생성 모듈(106)에 의해 임시 입자가 생성된 경우, 적대적 예제 생성 모듈(104)은 생성된 임시 입자에 대해 적합도를 산출할 수 있다. 적대적 예제 생성 모듈(104)은 생성된 임시 입자를 탐색 공간에서 반복적으로 이동시켜 적합도를 산출하고, 적합도를 갱신할 수 있다.
- [0149] 즉, 적대적 예제를 생성함에 있어 처음에 초기 입자를 많이 생성할수록 공격 성공률이 높아지나, 비정상적인 질의 수로 인해 공격 탐지 위험도 크게 증가하게 된다. 이에 개시되는 실시예에서는, 인공 신경망 모델(50)에 대한 질의 수가 기 설정된 임계 개수 이하로 발생하도록 처음에 초기 입자의 생성 개수를 제한할 수 있다. 그리고, 임시 입자 생성 이벤트가 발생하면, 임시 입자를 추가적으로 생성함으로써 공격 탐지 위험을 감소시키면서 적대적 예제가 되는 최적의 위치를 산출할 수 있게 된다.
- [0150] 적대적 예제 생성 모듈(104)은 적합도가 기 설정된 임계 적합도 이상이 되는 초기 입자 또는 임시 입자를 적대적 예제로 추출할 수 있다. 적대적 예제 생성 모듈(104)은 추출한 적대적 예제들을 적대적 예제 데이터 셋으로 하여 인공 신경망 모델(50)에 대해 적대적 공격을 수행할 수 있다.
- [0151] 개시되는 실시예에 의하면, 원본 데이터에 노이즈를 추가하여 초기 입자를 생성하고, 초기 입자를 탐색 공간 상에서 반복적으로 이동시키면서 적합도를 산출함으로써, 적대적 공격을 수행하기 위한 최적의 적대적 예제를 추출할 수 있게 된다. 그리고, 임시 입자 생성 이벤트가 발생하면, 임시 입자를 추가적으로 생성함으로써 공격 탐지 위험을 감소시킬 수 있고 효율적인 적대적 공격을 수행할 수 있게 된다.
- [0152] 본 명세서에서 모듈이라 함은, 본 발명의 기술적 사상을 수행하기 위한 하드웨어 및 상기 하드웨어를 구동하기 위한 소프트웨어의 기능적, 구조적 결합을 의미할 수 있다. 예컨대, 상기 "모듈"은 소정의 코드와 상기 소정의

코드가 수행되기 위한 하드웨어 리소스의 논리적인 단위를 의미할 수 있으며, 반드시 물리적으로 연결된 코드를 의미하거나, 한 종류의 하드웨어를 의미하는 것은 아니다.

- [0154] 도 5는 본 발명의 일 실시예에 따른 적대적 공격 방법을 나타낸 흐름도이고, 도 6은 본 발명의 일 실시예에 따른 적대적 공격 방법을 개략적으로 나타낸 도면이다. 도시된 흐름도에서는 상기 방법을 복수 개의 단계로 나누어 기재하였으나, 적어도 일부의 단계들은 순서를 바꾸어 수행되거나, 다른 단계와 결합되어 함께 수행되거나, 생략되거나, 세부 단계들로 나뉘어 수행되거나, 또는 도시되지 않은 하나 이상의 단계가 부가되어 수행될 수 있다.
- [0155] 도 5 및 도 6을 참조하면, 적대적 공격 장치(100)는 원본 데이터에 노이즈(또는 섭동)를 추가하여 초기 입자를 생성할 수 있다(S 101). 이때, 생성된 각 초기 입자들은 원본 데이터(예를 들어, 오디오 데이터 등)에 추가된 노이즈에 따라 탐색 공간 내에 각각 위치하게 된다. 적대적 공격 장치(100)는 초기 입자들의 처음 위치에서의 적합도를 각각 산출하고 그 위치 및 적합도를 저장할 수 있다.
- [0156] 다음으로, 적대적 공격 장치(100)는 탐색 공간 내 초기 입자들을 이동시키고, 이동된 위치에서 각 초기 입자들의 적합도를 산출할 수 있다(S 103). 적대적 공격 장치(100)는 초기 입자들이 이동됨에 따라 초기 입자들의 위치 및 적합도를 갱신하여 저장할 수 있다.
- [0157] 다음으로, 적대적 공격 장치(100)는 기 설정된 임시 입자 생성 이벤트가 발생하는지 확인할 수 있다(S 105). 예를 들어, 적대적 공격 장치(100)는 기 설정된 최대 반복 횟수까지 초기 입자를 이동시켰어도 적합도가 임계 적합도에 도달하지 못하거나 초기 입자의 이동을 일정 횟수 이상 반복적으로 수행하였는데도 적합도가 더 이상 개선되지 않는 경우 임시 입자 생성 이벤트가 발생한 것으로 판단할 수 있다.
- [0158] 단계 S 105의 확인 결과, 임시 입자 생성 이벤트가 발생하는 경우, 적대적 공격 장치(100)는 기 생성된 초기 입자에 기반하여 임시 입자를 추가로 생성할 수 있다(S 107). 적대적 공격 장치(100)는 생성한 임시 입자의 처음 위치에서의 적합도를 산출하고, 임시 입자들을 이동시키면서 적합도를 산출할 수 있다.
- [0159] 다음으로, 적대적 공격 장치(100)는 적합도가 기 설정된 임계 적합도 이상이 되는 초기 입자 또는 임시 입자를 적대적 예제로 추출할 수 있다(S 109). 여기서, 적대적 예제는 인공 신경망 모델(50)이 해당 적대적 예제를 오분류 하는 것을 목표로 생성된 것일 수 있다.
- [0160] 다음으로, 적대적 공격 장치(100)는 추출한 적대적 예제들을 적대적 예제 데이터 셋으로 하여 인공 신경망 모델(50)에 대해 적대적 공격을 수행할 수 있다(S 111). 적대적 공격 장치(100)는 적대적 공격의 수행 결과를 확인하여 리포트를 작성할 수 있다.
- [0162] 도 7은 예시적인 실시예들에서 사용되기에 적합한 컴퓨팅 장치를 포함하는 컴퓨팅 환경(10)을 예시하여 설명하기 위한 블록도이다. 도시된 실시예에서, 각 컴포넌트들은 이하에 기술된 것 이외에 상이한 기능 및 능력을 가질 수 있고, 이하에 기술된 것 이외에도 추가적인 컴포넌트를 포함할 수 있다.
- [0163] 도시된 컴퓨팅 환경(10)은 컴퓨팅 장치(12)를 포함한다. 일 실시예에서, 컴퓨팅 장치(12)는 적대적 공격 장치(100)일 수 있다.
- [0164] 컴퓨팅 장치(12)는 적어도 하나의 프로세서(14), 컴퓨터 판독 가능 저장 매체(16) 및 통신 버스(18)를 포함한다. 프로세서(14)는 컴퓨팅 장치(12)로 하여금 앞서 언급된 예시적인 실시예에 따라 동작하도록 할 수 있다. 예컨대, 프로세서(14)는 컴퓨터 판독 가능 저장 매체(16)에 저장된 하나 이상의 프로그램들을 실행할 수 있다. 상기 하나 이상의 프로그램들은 하나 이상의 컴퓨터 실행 가능 명령어를 포함할 수 있으며, 상기 컴퓨터 실행 가능 명령어는 프로세서(14)에 의해 실행되는 경우 컴퓨팅 장치(12)로 하여금 예시적인 실시예에 따른 동작들을 수행하도록 구성될 수 있다.
- [0165] 컴퓨터 판독 가능 저장 매체(16)는 컴퓨터 실행 가능 명령어 내지 프로그램 코드, 프로그램 데이터 및/또는 다른 적합한 형태의 정보를 저장하도록 구성된다. 컴퓨터 판독 가능 저장 매체(16)에 저장된 프로그램(20)은 프로세서(14)에 의해 실행 가능한 명령어의 집합을 포함한다. 일 실시예에서, 컴퓨터 판독 가능 저장 매체(16)는 메모리(랜덤 액세스 메모리와 같은 휘발성 메모리, 비휘발성 메모리, 또는 이들의 적절한 조합), 하나 이상의 자기 디스크 저장 디바이스들, 광학 디스크 저장 디바이스들, 플래시 메모리 디바이스들, 그 밖에 컴퓨팅 장치

(12)에 의해 액세스되고 원하는 정보를 저장할 수 있는 다른 형태의 저장 매체, 또는 이들의 적합한 조합일 수 있다.

[0166] 통신 버스(18)는 프로세서(14), 컴퓨터 판독 가능 저장 매체(16)를 포함하여 컴퓨팅 장치(12)의 다른 다양한 컴포넌트들을 상호 연결한다.

[0167] 컴퓨팅 장치(12)는 또한 하나 이상의 입출력 장치(24)를 위한 인터페이스를 제공하는 하나 이상의 입출력 인터페이스(22) 및 하나 이상의 네트워크 통신 인터페이스(26)를 포함할 수 있다. 입출력 인터페이스(22) 및 네트워크 통신 인터페이스(26)는 통신 버스(18)에 연결된다. 입출력 장치(24)는 입출력 인터페이스(22)를 통해 컴퓨팅 장치(12)의 다른 컴포넌트들에 연결될 수 있다. 예시적인 입출력 장치(24)는 포인팅 장치(마우스 또는 트랙패드 등), 키보드, 터치 입력 장치(터치패드 또는 터치스크린 등), 음성 또는 소리 입력 장치, 다양한 종류의 센서 장치 및/또는 촬영 장치와 같은 입력 장치, 및/또는 디스플레이 장치, 프린터, 스피커 및/또는 네트워크 카드와 같은 출력 장치를 포함할 수 있다. 예시적인 입출력 장치(24)는 컴퓨팅 장치(12)를 구성하는 일 컴포넌트로서 컴퓨팅 장치(12)의 내부에 포함될 수도 있고, 컴퓨팅 장치(12)와는 구별되는 별개의 장치로 컴퓨팅 장치(12)와 연결될 수도 있다.

[0169] 이상에서 본 발명의 대표적인 실시예들을 상세하게 설명하였으나, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자는 상술한 실시예에 대하여 본 발명의 범주에서 벗어나지 않는 한도 내에서 다양한 변형이 가능함을 이해할 것이다. 그러므로 본 발명의 권리범위는 설명된 실시예에 국한되어 정해져서는 안 되며, 후술하는 특허 청구범위뿐만 아니라 이 특허청구범위와 균등한 것들에 의해 정해져야 한다.

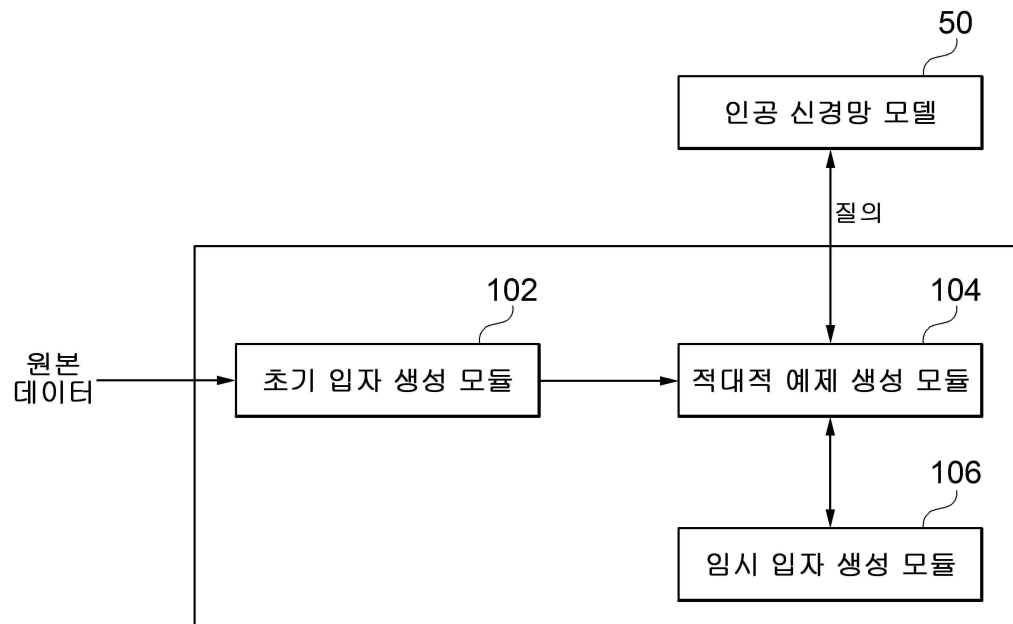
부호의 설명

[0171] 50 : 인공 신경망 모델
100 : 적대적 공격 장치
102 : 초기 입자 생성 모듈
104 : 적대적 예제 생성 모듈
106 : 임시 입자 생성 모듈

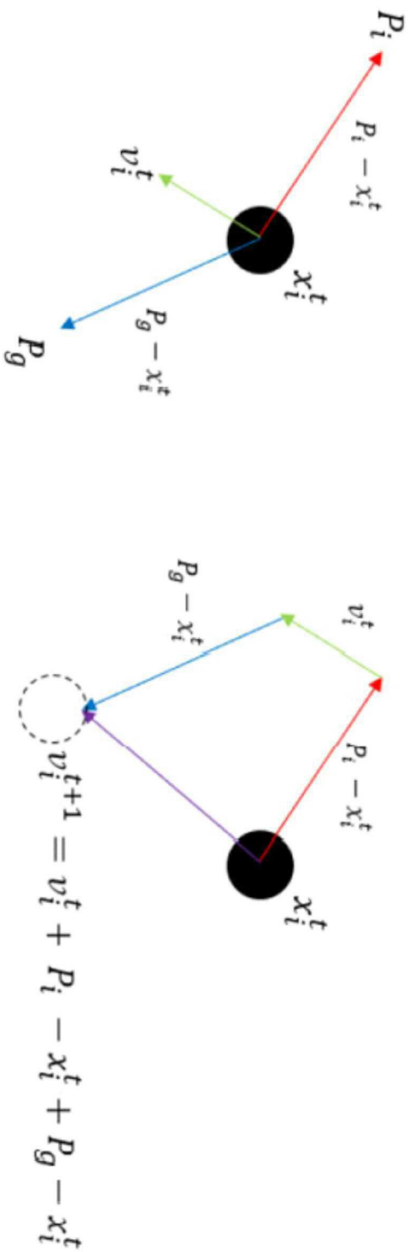
도면

도면1

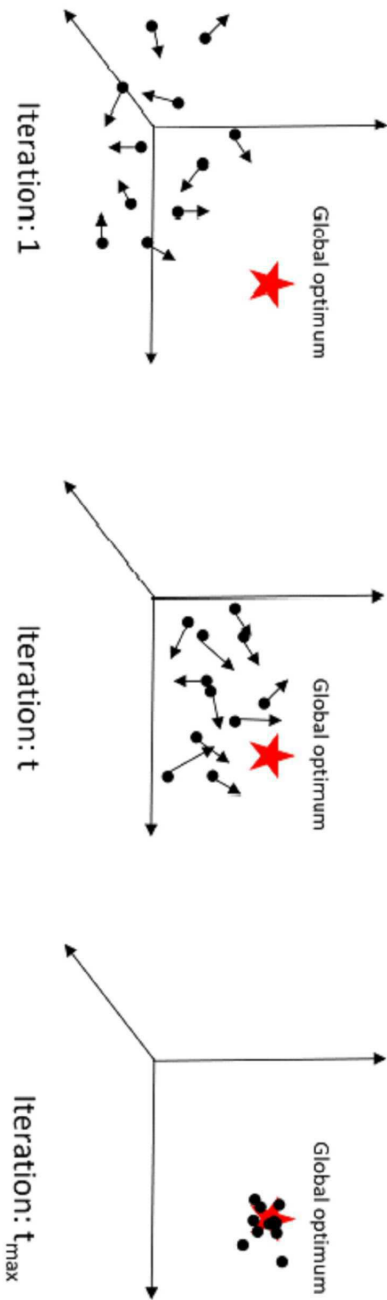
100



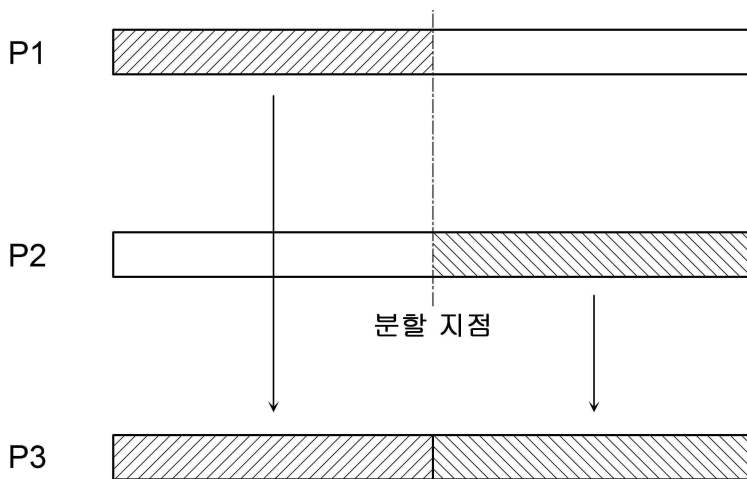
도면2



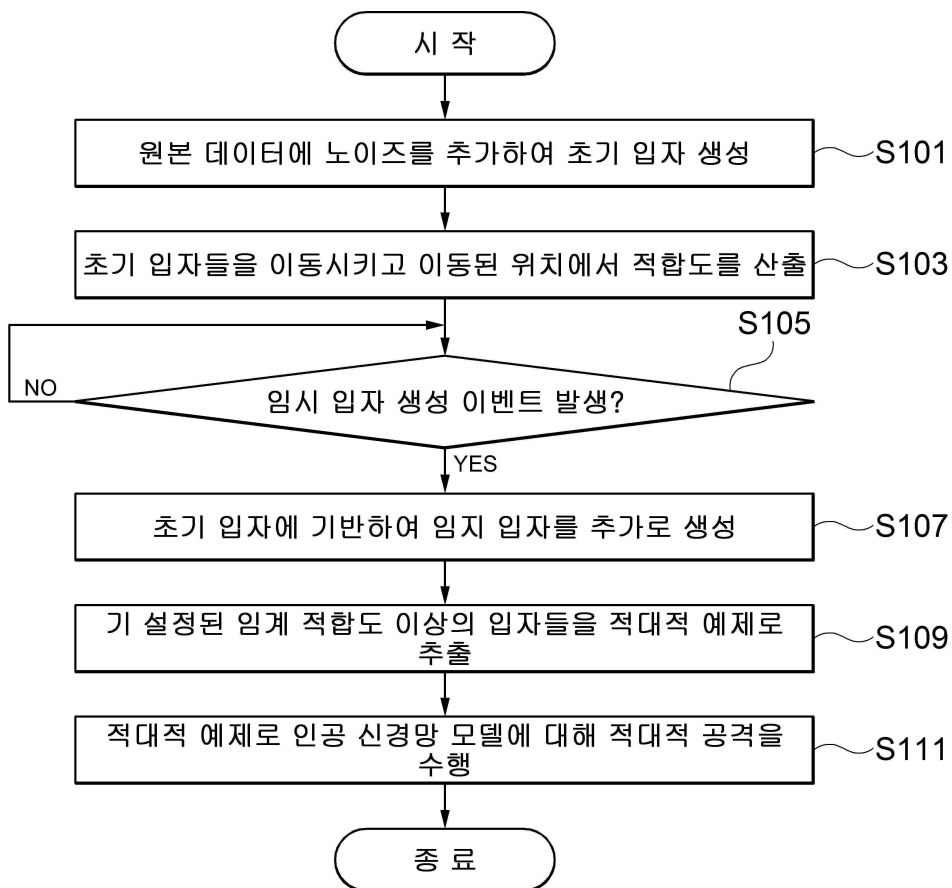
도면3



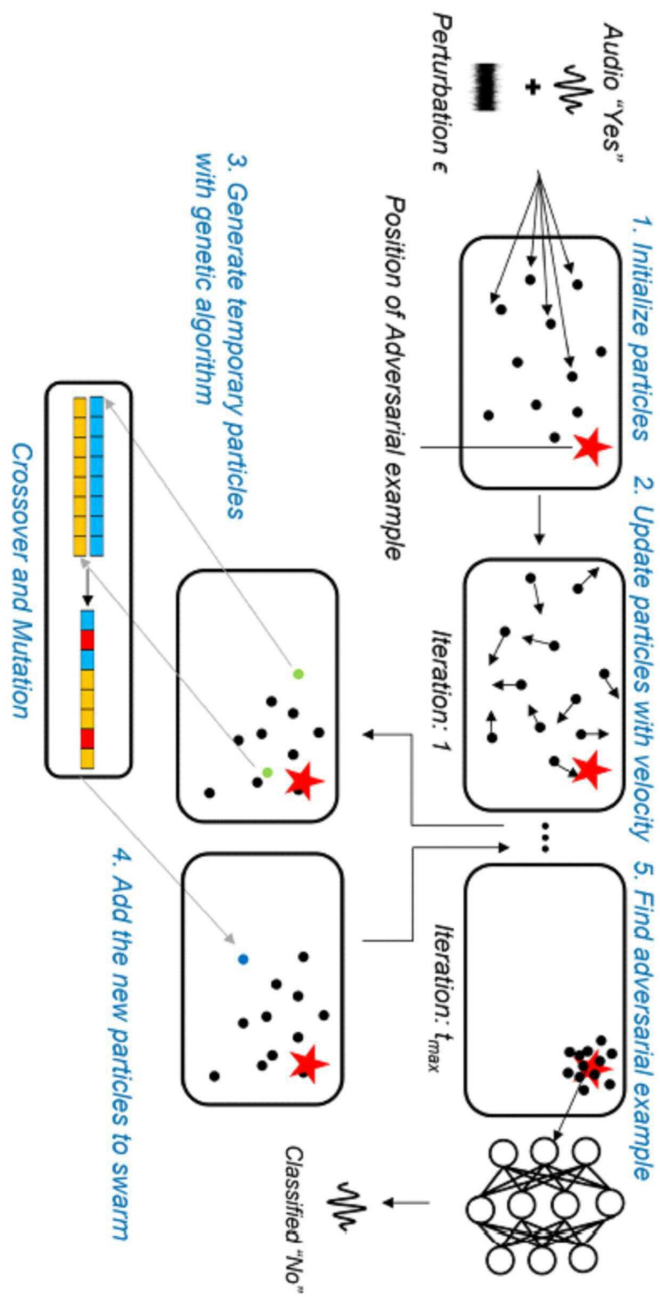
도면4



도면5



도면6



도면7

10

