(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0168539
(43) 공개일자 2023년12월14일

(51) 국제특허분류(Int. Cl.)
G06N 3/08 (2023.01) *G06N 3/04* (2023.01)
(52) CPC특허분류
G06N 3/08 (2023.01)
G06N 3/04 (2023.01)
(21) 출원번호 10-2022-0069150
(22) 출원일자 2022년06월07일
심사청구일자 2022년06월07일

(71) 출원인
세종대학교산학협력단
서울특별시 광진구 능동로 209 (군자동, 세종대학교)
(72) 발명자
윤주범
서울특별시 송파구 충민로4길 19, 704동 401호(장지동, 송파파인타운7단지)
서성관
서울특별시 광진구 면목로5길 30-3, 302호(군자동)
(뒷면에 계속)
(74) 대리인
특허법인티앤씨

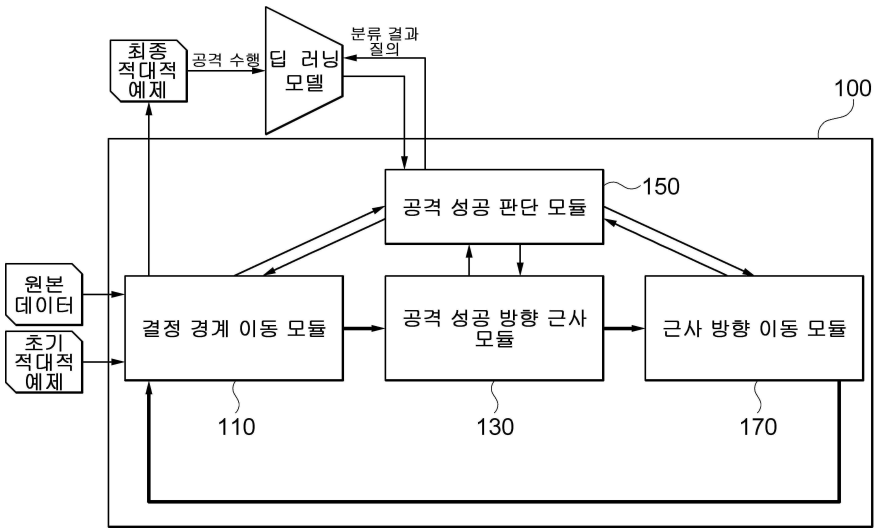
전체 청구항 수 : 총 19 항

(54) 발명의 명칭 **의사결정 기반 적대적 예제 생성 시스템 및 방법**

(57) 요 약

의사결정 기반 적대적 예제 생성 시스템 및 방법이 개시된다. 본 발명의 일 실시예에 따른 의사결정 기반 적대적 예제 생성 방법은, 제1 클래스로 분류되는 초기 적대적 예제와 제2 클래스로 분류되는 원본 데이터를 이용한 이진 탐색을 통해 상기 제1 클래스와 상기 제2 클래스에 대한 딥 러닝 모델의 결정 경계 상의 제1 적대적 예제를 획득하는 단계; 상기 딥 러닝 모델에 대한 적대적 공격 성공 여부에 기초하여 상기 제1 적대적 예제와 상기 원본 데이터 간에 직교하는 영역 내에서 상기 제1 적대적 예제에 대한 이동 방향을 결정하는 단계; 및 이동 방향으로 상기 제1 적대적 예제를 이동시켜 제2 적대적 예제를 생성하는 단계를 포함한다.

대 표 도



(72) 발명자

문현준

서울특별시 강북구 삼양로42길 15(미아동)

손배훈

서울특별시 광진구 광나루로13길 13, 304호(군자동)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711160322
과제번호	2020-0-01602-003
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신방송혁신인재양성
연구과제명	지능형 사이버 위협 대응 기술 개발 및 인력양성
기 여 율	1/1
과제수행기관명	숭실대학교 산학협력단
연구기간	2022.01.01 ~ 2022.12.31

명 세 서

청구범위

청구항 1

제1 클래스로 분류되는 초기 적대적 예제와 제2 클래스로 분류되는 원본 데이터를 이용한 이진 탐색을 통해 상기 제1 클래스와 상기 제2 클래스에 대한 딥 러닝 모델의 결정 경계 상의 제1 적대적 예제를 획득하는 단계;

상기 딥 러닝 모델에 대한 적대적 공격 성공 여부에 기초하여 상기 제1 적대적 예제와 상기 원본 데이터 간에 직교하는 영역 내에서 상기 제1 적대적 예제에 대한 이동 방향을 결정하는 단계; 및

상기 이동 방향으로 상기 제1 적대적 예제를 이동시켜 제2 적대적 예제를 생성하는 단계를 포함하는, 의사결정 기반 적대적 예제 생성 방법.

청구항 2

청구항 1에 있어서,

상기 이동 방향을 결정하는 단계는, 상기 직교하는 영역 내 복수의 방향에 대응하는 복수의 방향 벡터를 생성하는 단계;

상기 복수의 방향 벡터 각각과 상기 제1 적대적 예제에 기초하여 상기 복수의 방향 각각에 대한 상기 적대적 공격 성공 여부를 판단하는 단계;

상기 적대적 공격 성공 여부에 기초하여 상기 복수의 방향 벡터 각각에 대한 가중치를 결정하는 단계; 및

상기 가중치 및 상기 복수의 방향 벡터에 기초하여 상기 이동 방향을 결정하는 단계를 포함하는, 의사결정 기반 적대적 예제 생성 방법.

청구항 3

청구항 2에 있어서,

상기 가중치 및 상기 복수의 방향 벡터에 기초하여 상기 이동 방향을 결정하는 단계는, 상기 가중치에 기초한 상기 복수의 방향 벡터에 대한 가중 평균을 이용하여 상기 이동 방향에 대응하는 방향 벡터를 결정하는, 의사결정 기반 적대적 예제 생성 방법.

청구항 4

청구항 3에 있어서,

상기 가중치를 결정하는 단계는, 상기 복수의 방향 중 상기 적대적 공격이 성공한 방향에 대응하는 방향 벡터에 대한 가중치를 1로 결정하고, 상기 복수의 방향 중 상기 적대적 공격이 실패한 방향에 대응하는 방향 벡터에 대한 가중치를 -1로 결정하는, 의사결정 기반 적대적 예제 생성 방법.

청구항 5

청구항 2에 있어서,

상기 복수의 방향 벡터를 생성하는 단계는,

수학식 1을 기초로 상기 복수의 방향 벡터 각각을 생성하고,

(수학식 1)

$$u = u - proj_{x_t - x^*} u$$

상기 u는 방향 벡터, x_t는 상기 제1 적대적 예제, x^*는 상기 원본 데이터인, 의사결정 기반 적대적 예제 생성 방법.

청구항 6

청구항 1에 있어서,

상기 제2 적대적 예제를 생성하는 단계는, 상기 이동 방향으로 상기 제1 적대적 예제를 기 설정된 크기만큼 이동시키되, 상기 딥 러닝 모델에 대한 적대적 공격이 성공할 때까지 이동 거리 변경 및 이동을 반복 수행하여 상기 제2 적대적 예제를 결정하는, 의사결정 기반 적대적 예제 생성 방법.

청구항 7

청구항 6에 있어서,

상기 공격이 성공할 때까지 이동 거리 변경 및 이동을 반복 수행하는 단계는,

상기 제1 적대적 예제를 기 설정된 크기만큼 이동시키는 단계;

상기 이동에 따른 공격이 성공했는지 여부를 확인하는 단계; 및

확인 결과 공격이 실패한 경우, 이동 거리를 절반으로 줄여 상기 제1 적대적 예제를 이동시키는 단계를 포함하는, 의사결정 기반 적대적 예제 생성 방법.

청구항 8

청구항 1에 있어서,

기 설정된 종료 조건을 만족할 때까지 상기 제2 적대적 예제를 상기 초기 적대적 예제로 이용하여 상기 획득하는 단계 내지 상기 생성하는 단계를 반복 수행하는 단계를 더 포함하는, 의사결정 기반 적대적 공격 방법.

청구항 9

청구항 1에 있어서,

상기 제2 적대적 예제를 생성하는 단계 이후에,

상기 원본 데이터에 인접한 결정 경계 상에 위치하는 최종 적대적 예제를 결정하는 단계를 더 포함하는, 의사결정 기반 적대적 공격 방법.

청구항 10

청구항 9에 있어서,

상기 최종 적대적 예제를 결정하는 단계 이후에,

상기 최종 적대적 예제를 이용하여 딥 러닝 모델로 공격을 수행하는 단계를 더 포함하는, 의사결정 기반 적대적 예제 생성 방법.

청구항 11

제1 클래스로 분류되는 초기 적대적 예제와 제2 클래스로 분류되는 원본 데이터를 이용한 이진 탐색을 통해 상기 제1 클래스와 상기 제2 클래스에 대한 딥 러닝 모델의 결정 경계 상의 제1 적대적 예제를 획득하기 위한 결정 경계 이동 모듈;

상기 딥 러닝 모델에 대한 적대적 공격 성공 여부에 기초하여 상기 제1 적대적 예제와 상기 원본 데이터 간에 직교하는 영역 내에서 상기 제1 적대적 예제에 대한 이동 방향을 결정하기 위한 공격 성공 방향 근사 모듈;

상기 딥 러닝 모델로 분류 결과를 질의하여 적대적 공격 성공 여부를 판단하기 위한 공격 성공 판단 모듈; 및

상기 이동 방향으로 상기 제1 적대적 예제를 이동시켜 제2 적대적 예제를 생성하기 위한 근사 방향 이동 모듈을 포함하는, 의사결정 기반 적대적 예제 생성 시스템.

청구항 12

청구항 11에 있어서,

상기 공격 성공 방향 근사 모듈은, 상기 직교하는 영역 내 복수의 방향에 대응하는 복수의 방향 벡터를 생성하고, 상기 복수의 방향 벡터 각각과 상기 제1 적대적 예제에 기초하여 상기 복수의 방향 각각에 대한 상기 적대적 공격 성공 여부를 판단하며, 상기 적대적 공격 성공 여부에 기초하여 상기 복수의 방향 벡터 각각에 대한 가중치를 결정하고, 및 상기 가중치 및 상기 복수의 방향 벡터에 기초하여 상기 이동 방향을 결정하는, 의사결정 기반 적대적 예제 생성 시스템.

청구항 13

청구항 12에 있어서,

상기 공격 성공 방향 근사 모듈은, 상기 가중치에 기초한 상기 복수의 방향 벡터에 대한 가중 평균을 이용하여 상기 이동 방향에 대응하는 방향 벡터를 결정하는, 의사결정 기반 적대적 예제 생성 시스템.

청구항 14

청구항 13에 있어서,

상기 공격 성공 방향 근사 모듈은, 상기 복수의 방향 중 상기 적대적 공격이 성공한 방향에 대응하는 방향 벡터에 대한 가중치를 1로 결정하고, 상기 복수의 방향 중 상기 적대적 공격이 실패한 방향에 대응하는 방향 벡터에 대한 가중치를 -1로 결정하는, 의사결정 기반 적대적 예제 생성 시스템.

청구항 15

청구항 12에 있어서,

상기 공격 성공 방향 근사 모듈은,

수학식 1을 기초로 상기 복수의 방향 벡터 각각을 생성하고,

(수학식 1)

$$u = u - proj_{x_t - x^*} u$$

상기 u는 방향 벡터, x_t는 상기 제1 적대적 예제, x^*는 상기 원본 데이터인, 의사결정 기반 적대적 예제 생성 시스템.

청구항 16

청구항 11에 있어서,

상기 근사 방향 이동 모듈은,

상기 이동 방향으로 상기 제1 적대적 예제를 기 설정된 크기만큼 이동시키되, 상기 딥 러닝 모델에 대한 적대적 공격이 성공할 때까지 이동 거리 변경 및 이동을 반복 수행하여 상기 제2 적대적 예제를 결정하는, 의사결정 기반 적대적 예제 생성 시스템.

청구항 17

청구항 16에 있어서,

상기 근사 방향 이동 모듈은,

상기 제1 적대적 예제를 기 설정된 크기만큼 이동시키고, 상기 이동에 따른 공격이 성공했는지 여부를 확인하여, 확인 결과 공격이 실패한 경우 이동 거리를 절반으로 줄여 상기 제1 적대적 예제를 이동시키는, 의사결정 기반 적대적 예제 생성 시스템.

청구항 18

청구항 11에 있어서,

상기 결정 경계 이동 모듈은, 상기 원본 데이터에 인접한 결정 경계 상에 위치하는 최종 적대적 예제를 결정하는 것을 더 포함하는, 의사결정 기반 적대적 공격 시스템.

청구항 19

청구항 18에 있어서,

상기 결정 경계 이동 모듈은,

상기 최종 적대적 예제를 이용하여 딥 러닝 모델로 공격을 수행하는, 의사결정 기반 적대적 예제 생성 시스템.

발명의 설명

기 술 분 야

[0001] 개시되는 실시예들은 의사결정 기반 적대적 예제 생성 시스템 및 방법과 관련된다.

배 경 기 술

[0002] 적대적 공격은 원본 데이터에 사람이 인식할 수 없는 작은 노이즈를 더하여 딥 러닝 모델의 출력결과를 공격자가 원하는 결과로 바꾸는 공격 방법이다.

[0003] 일반적으로, 상술한 적대적 공격은 딥 러닝 모델의 훈련 과정과 유사하게 동작하는데, 공격 성공을 위한 목적함수를 정의하고 목적함수를 최소화 또는 최대화하기 위해 정의한 목적함수의 기울기를 원본 데이터에 더할 수 있다. 즉, 공격자가 원하는 분류 결과를 낼 수 있도록 반복적으로 목적 함수의 기울기인 노이즈를 원본 데이터에 추가하는 방식으로 동작할 수 있다. 상술한 적대적 공격은 공격자에게 주어지는 정보에 따라 화이트 박스 공격과 블랙박스 공격으로 구분될 수 있다.

[0004] 한편, AI방식을 이용하여 원본 이미지로부터 특징을 추출하고 추출된 특징을 파괴시킨 새로운 적대적 이미지를 생성하여 지능형 봇(Bot)에서 사용되는 딥러닝 방식의 객체 인식이 올바르게 동작하지 못하도록 적대적 이미지를 생성하는 기술이 제안되었다.

선행기술문헌

특허문헌

[0006] (특허문헌 0001) 대한민국 등록특허공보 제10-2253402호 (2021. 05. 12.)

발명의 내용

해결하려는 과제

[0007] 개시된 실시예들은 의사결정 기반 공격기법의 딥 러닝 모델로의 질의를 줄이기 위한 의사결정 기반 적대적 예제 생성 시스템 및 방법을 제공하고자 한다.

과제의 해결 수단

[0009] 일 실시예에 따른 의사결정 기반 적대적 예제 생성 방법은, 제1 클래스로 분류되는 초기 적대적 예제와 제2 클래스로 분류되는 원본 데이터를 이용한 이진 탐색을 통해 상기 제1 클래스와 상기 제2 클래스에 대한 딥 러닝 모델의 결정 경계 상의 제1 적대적 예제를 획득하는 단계; 상기 딥 러닝 모델에 대한 적대적 공격 성공 여부에 기초하여 상기 제1 적대적 예제와 상기 원본 데이터 간에 직교하는 영역 내에서 상기 제1 적대적 예제에 대한 이동 방향을 결정하는 단계; 및 상기 이동 방향으로 상기 제1 적대적 예제를 이동시켜 제2 적대적 예제를 생성하는 단계를 포함한다.

[0010] 또한, 상기 이동 방향을 결정하는 단계는, 상기 직교하는 영역 내 복수의 방향에 대응하는 복수의 방향 벡터를 생성하는 단계; 상기 복수의 방향 벡터 각각과 상기 제1 적대적 예제에 기초하여 상기 복수의 방향 각각에 대한 상기 적대적 공격 성공 여부를 판단하는 단계; 상기 적대적 공격 성공 여부에 기초하여 상기 복수의 방향 벡터 각각에 대한 가중치를 결정하는 단계; 및 상기 가중치 및 상기 복수의 방향 벡터에 기초하여 상기 이동 방향을 결정하는 단계를 포함할 수 있다.

[0011] 또한, 상기 가중치 및 상기 복수의 방향 벡터에 기초하여 상기 이동 방향을 결정하는 단계는, 상기 가중치에 기초한 상기 복수의 방향 벡터에 대한 가중 평균을 이용하여 상기 이동 방향에 대응하는 방향 벡터를 결정할 수 있다.

[0012] 또한, 상기 가중치를 결정하는 단계는, 상기 복수의 방향 중 상기 적대적 공격이 성공한 방향에 대응하는 방향 벡터에 대한 가중치를 1로 결정하고, 상기 복수의 방향 중 상기 적대적 공격이 실패한 방향에 대응하는 방향 벡터에 대한 가중치를 -1로 결정할 수 있다.

[0013] 또한, 상기 복수의 방향 벡터를 생성하는 단계는, 수학식 1을 기초로 상기 복수의 방향 벡터 각각을 생성하고,

[0014] (수학식 1)

[0015] $$u = u - proj_{x_t - x^*} u$$

[0016] 상기 u는 방향 벡터, 상기 x_t는 상기 제1 적대적 예제, 상기 x^*는 상기 원본 데이터일 수 있다.

[0017] 또한, 상기 제2 적대적 예제를 생성하는 단계는, 상기 이동 방향으로 상기 제1 적대적 예제를 기 설정된 크기만큼 이동시키되, 상기 딥 러닝 모델에 대한 적대적 공격이 성공할 때까지 이동 거리 변경 및 이동을 반복 수행하여 상기 제2 적대적 예제를 결정할 수 있다.

[0018] 또한, 상기 공격이 성공할 때까지 이동 거리 변경 및 이동을 반복 수행하는 단계는, 상기 제1 적대적 예제를 기 설정된 크기만큼 이동시키는 단계; 상기 이동에 따른 공격이 성공했는지 여부를 확인하는 단계; 및 확인 결과 공격이 실패한 경우, 이동 거리를 절반으로 줄여 상기 제1 적대적 예제를 이동시키는 단계를 포함할 수 있다.

[0019] 또한, 상기 의사결정 기반 적대적 공격 방법은, 기 설정된 종료 조건을 만족할 때까지 상기 제2 적대적 예제를 상기 초기 적대적 예제로 이용하여 상기 획득하는 단계 내지 상기 생성하는 단계를 반복 수행하는 단계를 더 포함할 수 있다.

[0020] 또한, 상기 의사결정 기반 적대적 공격 방법은, 상기 제2 적대적 예제를 생성하는 단계 이후에, 상기 원본 데이터에 인접한 결정 경계 상에 위치하는 최종 적대적 예제를 결정하는 단계를 더 포함할 수 있다.

[0021] 또한, 상기 의사결정 기반 적대적 공격 방법은, 상기 최종 적대적 예제를 결정하는 단계 이후에, 상기 최종 적대적 예제를 이용하여 딥 러닝 모델로 공격을 수행하는 단계를 더 포함할 수 있다.

[0022] 다른 실시예에 따른 의사결정 기반 적대적 예제 생성 시스템은, 제1 클래스로 분류되는 초기 적대적 예제와 제2 클래스로 분류되는 원본 데이터를 이용한 이진 탐색을 통해 상기 제1 클래스와 상기 제2 클래스에 대한 딥 러닝 모델의 결정 경계 상의 제1 적대적 예제를 획득하기 위한 결정 경계 이동 모듈; 상기 딥 러닝 모델에 대한 적대적 공격 성공 여부에 기초하여 상기 제1 적대적 예제와 상기 원본 데이터 간에 직교하는 영역 내에서 상기 제1 적대적 예제에 대한 이동 방향을 결정하기 위한 공격 성공 방향 근사 모듈; 상기 딥 러닝 모델로 분류 결과를 질의하여 적대적 공격 성공 여부를 판단하기 위한 공격 성공 판단 모듈; 및 상기 이동 방향으로 상기 제1 적대적 예제를 이동시켜 제2 적대적 예제를 생성하기 위한 근사 방향 이동 모듈을 포함한다.

[0023] 또한, 상기 공격 성공 방향 근사 모듈은, 상기 직교하는 영역 내 복수의 방향에 대응하는 복수의 방향 벡터를 생성하고, 상기 복수의 방향 벡터 각각과 상기 제1 적대적 예제에 기초하여 상기 복수의 방향 각각에 대한 상기 적대적 공격 성공 여부를 판단하며, 상기 적대적 공격 성공 여부에 기초하여 상기 복수의 방향 벡터 각각에 대한 가중치를 결정하고, 및 상기 가중치 및 상기 복수의 방향 벡터에 기초하여 상기 이동 방향을 결정할 수 있다.

[0024] 또한, 상기 공격 성공 방향 근사 모듈은, 상기 가중치에 기초한 상기 복수의 방향 벡터에 대한 가중 평균을 이용하여 상기 이동 방향에 대응하는 방향 벡터를 결정할 수 있다.

[0025] 또한, 상기 공격 성공 방향 근사 모듈은, 상기 복수의 방향 중 상기 적대적 공격이 성공한 방향에 대응하는 방향 벡터에 대한 가중치를 1로 결정하고, 상기 복수의 방향 중 상기 적대적 공격이 실패한 방향에 대응하는 방향 벡터에 대한 가중치를 -1로 결정할 수 있다.

[0026] 또한, 상기 공격 성공 방향 근사 모듈은, 수학식 1을 기초로 상기 복수의 방향 벡터 각각을 생성하고,

[0027] (수학식 1)

[0028] $$u = u - proj_{x_t - x^*} u$$

[0029] 상기 u는 방향 벡터, 상기 x_t는 상기 제1 적대적 예제, 상기 x^*는 상기 원본 데이터일 수 있다.

[0030] 또한, 상기 근사 방향 이동 모듈은, 상기 이동 방향으로 상기 제1 적대적 예제를 기 설정된 크기만큼 이동시키되, 상기 딥 러닝 모델에 대한 적대적 공격이 성공할 때까지 이동 거리 변경 및 이동을 반복 수행하여 상기 제2 적대적 예제를 결정할 수 있다.

[0031] 또한, 상기 근사 방향 이동 모듈은, 상기 제1 적대적 예제를 기 설정된 크기만큼 이동시키고, 상기 이동에 따른 공격이 성공했는지 여부를 확인하여, 확인 결과 공격이 실패한 경우 이동 거리를 절반으로 줄여 상기 제1 적대적 예제를 이동시킬 수 있다.

[0032] 또한, 상기 결정 경계 이동 모듈은, 상기 원본 데이터에 인접한 결정 경계 상에 위치하는 최종 적대적 예제를 결정하는 것을 더 포함할 수 있다.

[0033] 또한, 상기 결정 경계 이동 모듈은, 상기 최종 적대적 예제를 이용하여 딥 러닝 모델로 공격을 수행할 수 있다.

발명의 효과

[0034] 개시되는 실시예들에 따르면, 딥 러닝 모델에 대한 적대적 공격을 위한 적대적 예제 생성 시 결정 경계 상으로 이동된 적대적 예제를 이동시킬 방향 벡터를 찾을 때 복수의 랜덤 벡터를 이용하되 전체 영역이 아닌 적대적 예제와 원본 데이터에 대해 직교하는 영역 내에 있는 랜덤 벡터를 이용함으로써 공격 성공 여부에 대한 딥 러닝 모델로의 질의 횟수를 현저히 감소시킬 수 있다는 효과를 기대할 수 있다.

[0035] 또한, 실시예들에 따르면, 의사결정 기반 공격기법의 기울기 예측 시 이용되는 방향 벡터 탐색 영역을 축소하기 때문에, 질의 효율성을 증대 시키고 실제 환경에서 적용될 수 있는 적대적 공격을 수행할 수 있다는 것이다.

도면의 간단한 설명

[0037] 도 1은 일 실시예에 따른 의사결정 기반 적대적 예제 생성 시스템을 설명하기 위한 블록도

도 2 내지 도 5는 일 실시예에 따른 의사결정 기반 적대적 예제를 산출하는 방법을 설명하기 위한 예시도

도 6은 일 실시예에 따른 의사결정 기반 적대적 예제 생성 방법을 설명하기 위한 예시도

도 7은 일 실시예에 따른 의사결정 기반 적대적 예제 생성 방법을 설명하기 위한 흐름도

도 8 내지 도 11은 도 7의 일부를 상세하게 설명하기 위한 흐름도

도 12는 일 실시예에 따른 컴퓨팅 장치를 포함하는 컴퓨팅 환경을 예시하여 설명하기 위한 블록도

발명을 실시하기 위한 구체적인 내용

[0038] 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.

[0039] 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.

[0040] 본 실시예는 블랙박스 공격기법에 관한 것으로서, 블랙박스 공격은 공격자에게 주어지는 정보에 따라 분류될 수 있다. 구체적으로, 공격자가 딥 러닝 모델의 분류 점수를 인지하고 있는 점수 기반 공격기법과 공격자가 오직 딥 러닝 모델의 최종 출력결과만을 알 수 있는 의사결정 기반 공격기법을 포함한다. 예를 들어, 개, 고양이 이미지를 분류하는 경우, 딥 러닝 모델의 분류 점수는 몇 퍼센트로 개 혹은 고양이로 분류하는지를 의미할 수 있다. 또한, 개, 고양이 이미지를 분류하는 경우, 딥 러닝 모델의 최종 출력결과는 최종 분류인 개 혹은 고양이일 수 있다.

[0041] 본 실시예에서 개시하는 블랙박스 공격기법은 의사결정 공격기법으로, 딥 러닝 모델의 최종 분류 결과만을 이용하여 적대적 예제로 시작하여 원본 데이터에 근접하도록 하는 방식으로 동작할 수 있다. 이에 대한 상세 설명은 후술하기로 한다.

[0042] 도 1은 일 실시예에 따른 의사결정 기반 적대적 예제 생성 시스템을 설명하기 위한 블록도이다.

[0043] 이하에서는, 일 실시예에 따른 의사결정 기반 적대적 예제를 산출하는 방법을 설명하기 위한 예시도인 도 2 내지 도 5, 일 실시예에 따른 의사결정 기반 적대적 예제 생성 방법을 설명하기 위한 예시도인 도 6을 참고하여 설명하기로 한다.

[0044] 도 1을 참고하면, 의사결정 기반 적대적 예제 생성 시스템(100)(이하, '적대적 예제 생성 시스템'이라 하기로 함)은 딥 러닝 모델의 분류 결과를 이용하는 의사 결정 기반 시스템으로서, 결정 경계 이동 모듈(110), 공격 성

공 방향 근사 모듈(130), 공격 성공 판단 모듈(150) 및 근사 방향 이동 모듈(170)을 포함한다.

[0045] 결정 경계 이동 모듈(110)은 적대적 예제를 결정 경계 상으로 이동시키는 구성일 수 있다.

[0046] 보다 상세히 설명하면, 결정 경계 이동 모듈(110)은 제1 클래스로 분류되는 초기 적대적 예제와 제2 클래스로 분류되는 원본 데이터를 이용한 이진 탐색을 통해 제1 클래스와 제2 클래스에 대한 딥 러닝 모델의 결정 경계 상의 제1 적대적 예제를 획득할 수 있다. 이때, 제1 클래스와 같이 공격 목표 클래스로 분류되는 데이터를 초기 적대적 예제로 설정할 수 있다.

[0047] 결정 경계 이동 모듈(110)은 초기 적대적 예제가 결정 경계 상에 위치하는지 여부를 확인하고, 확인 결과 결정 경계 상에 미 위치하는 경우, 이진 탐색을 통해 결정 경계 상에 위치하는 제1 적대적 예제를 획득하는 것을 반복 수행할 수 있다.

[0048] 구체적으로, 결정 경계 이동 모듈(110)은 공격 성공 판단 모듈(150)의 질의 결과를 기초로 초기 적대적 예제와 원본 데이터 사이의 결정 경계를 이진 탐색을 이용하여 탐색하고, 초기 적대적 예제를 탐색된 결정 경계상의 제1 적대적 예제로 업데이트할 수 있다. 즉, 도 2에서 도시하는 바와 같이, 결정 경계 이동 모듈(110)은 이진 탐색을 통해 초기 적대적 예제를 결정 경계 상으로 이동시켜 제1 적대적 예제(x_t)를 획득할 수 있다.

[0049] 이때, 제1 적대적 예제는 초기 적대적 예제를 이동시키면서 최종 적대적 예제(예를 들어, 도 5의 x_{t+2})를 획득하는 과정에서 발생하는 결정 경계 상의 적대적 예제를 의미하는 것으로서, 초기 적대적 예제 및 최종 적대적 예제와 구분하기 위해 명명한 것일 수 있다.

[0050] 결정 경계 이동 모듈(110)은 원본 데이터에 인접한 결정 경계 상에 위치하는 최종 적대적 예제를 결정할 수 있다.

[0051] 구체적으로, 도 5를 참고하면, 결정 경계 이동 모듈(110)은 제1 적대적 예제가 원본 데이터에 인접한 결정 경계 상에 위치하는지 여부를 확인하고, 확인 결과 결정 경계 상에 미 위치하는 경우 제1 적대적 예제와 원본 데이터 간의 이진 탐색을 수행하여 원본 데이터에 인접한 결정 경계 상에 위치하는 최종 적대적 예제(x_{t+2})를 획득할 수 있다.

[0052] 결정 경계 이동 모듈(110)은 결정 경계 상에 위치하되, 공격 목표 클래스 분류를 유지하고 원본 데이터와의 차이가 기 설정된 기준 이하로 작은 최종 적대적 예제를 산출할 수 있다. 즉, 본 실시예는 공격 목표 클래스로 분류되는 데이터를 초기 적대적 예제로 설정하여 분류는 유지하되 원본 데이터와의 차이를 점차 줄이는 방식으로 적대적 예제를 수정하여 원본 데이터에 인접한 결정 경계 상의 최종 적대적 예제를 산출하는 것이다.

[0053] 도 1을 참고하면, 결정 경계 이동 모듈(110)은 최종 적대적 예제를 이용하여 딥 러닝 모델로 공격을 수행할 수 있다. 예를 들어, 딥 러닝 모델은 DNN(Deep Neural Network) 모델일 수 있다.

[0054] 공격 성공 방향 근사 모듈(130)은 딥 러닝 모델에 대한 적대적 공격 성공 여부에 기초하여 제1 적대적 예제와 원본 데이터 간에 직교하는 영역(도 3의 A) 내에서 제1 적대적 예제에 대한 이동 방향을 결정할 수 있다.

[0055] 구체적으로, 공격 성공 방향 근사 모듈(130)은 직교하는 영역 내 복수의 방향에 대응하는 복수의 방향 벡터를 생성하고, 복수의 방향 벡터 각각과 제1 적대적 예제에 기초하여 복수의 방향 각각에 대한 적대적 공격 성공 여부를 판단하며, 적대적 공격 성공 여부에 기초하여 복수의 방향 벡터 각각에 대한 가중치를 결정하고, 및 가중치 및 복수의 방향 벡터에 기초하여 이동 방향을 결정할 수 있다.

[0056] 도 3과 같이, 공격 성공 방향 근사 모듈(130)은 $x_t - x_{orig}$ 와 직교하는 범위(A) 내에서 방향 벡터를 선택하므로 전체 공간에서 방향 벡터를 추출하는 일반적인 방식에 비해 질의 수를 줄일 수 있는 것이다. 이때, 전체 공간은 x_t를 기준으로 전 방향(예를 들어, 구 형태의 전 방향)을 의미하는 것일 수 있다. 즉, 공격 성공 방향 근사 모듈(130)은 원본 데이터로부터 시작하여 제1 적대적 예제에서 종료하는 벡터와 직교하는 영역(A) 내에서 적어도 하나 이상의 방향 벡터(V)를 산출할 수 있다.

[0057] 공격 성공 방향 근사 모듈(130)은 가중치에 기초한 복수의 방향 벡터에 대한 가중 평균을 이용하여 이동 방향에 대응하는 방향 벡터를 결정할 수 있다.

[0058] 공격 성공 방향 근사 모듈(130)은 복수의 방향 중 적대적 공격이 성공한 방향에 대응하는 방향 벡터에 대한 가중치를 1로 결정하고, 복수의 방향 중 적대적 공격이 실패한 방향에 대응하는 방향 벡터에 대한 가중치를 -1로 결정할 수 있다.

[0059] 구체적으로, 공격 성공 방향 근사 모듈(130)은 수학식 1을 기초로 복수의 방향 벡터 각각을 생성할 수 있다. 이때, 공격 성공 방향 근사 모듈(130)은 방향 벡터 u를 선택한 다음 수학식 1과 같이 $x_t - x_{orig}$와 수직이 되도록 처리할 수 있다.

[0060] (수학식 1)

[0061] $$u = u - proj_{x_t - x^*} u$$

[0062] 상기 u는 방향 벡터, x_t는 제1 적대적 예제, x^*는 원본 데이터일 수 있다.

[0063] 공격 성공 방향 근사 모듈(130)은 수학식 2를 이용하여 방향 벡터에 가중치를 적용할 수 있다. 이때, 공격 성공 방향 근사 모듈(130)은 제1 적대적 예제(x_t)가 결정 경계 상에 위치한다는 전제하게 수학식 2와 같이 몬테카를로 알고리즘을 이용하여 A_{x^*}를 근사할 수 있다.

[0064] (수학식 2)

[0065] $$A(x_t, \delta) = \frac{1}{B} \sum_{b=1}^{B} \phi_{x^*}(x_t + \delta u_b) u_b$$

[0066] 이때, x_t는 제1 적대적 예제, δ는 기 설정된 값 이하로 작은 크기의 파라미터 값, B는 벡터의 개수, u_b는 랜덤한 값일 수 있다.

[0067] 도 1에서 도시하는 바와 같이, 공격 성공 판단 모듈(150)은 결정 경계 이동 모듈(110), 공격 성공 방향 근사 모듈(130) 및 근사 방향 이동 모듈(170)과 연결되어, 딥 러닝 모델로 분류 결과를 질의하여 적대적 공격 성공 여부를 판단하기 위한 구성일 수 있다.

[0068] 공격 성공 판단 모듈(150)은 직교하는 영역 내에서 추출한 방향 벡터를 제1 적대적 예제에 합산하고 딥 러닝 모델로 질의하여 공격 성공 여부를 판단할 수 있다.

[0069] 근사 방향 이동 모듈(170)은 공격 성공 방향 근사 모듈(130)에 의해서 결정된 이동 방향으로 제1 적대적 예제를 이동시켜 제2 적대적 예제를 생성하기 위한 구성일 수 있다.

[0070] 구체적으로, 도 4를 참고하면, 근사 방향 이동 모듈(170)은 결정된 이동 방향으로 제1 적대적 예제(x_t)를 기 설정된 크기(S1)만큼 이동시키되, 딥 러닝 모델에 대한 적대적 공격이 성공할 때까지 이동 거리 변경 및 이동을 반복 수행하여 제2 적대적 예제를 결정할 수 있다.

[0071] 이때, 근사 방향 이동 모듈(170)은 제1 적대적 예제를 기 설정된 크기만큼 이동시키고, 이동에 따른 공격이 성공했는지 여부를 확인하여, 확인 결과 공격이 실패한 경우 이동 거리를 절반으로 줄여 제1 적대적 예제를 이동시킬 수 있다. 적대적 예제 생성 시스템(200)은 후술하는 수학식 3을 기초로 제1 적대적 예제를 너무 많이 이동하여 공격이 실패하는 경우 공격이 성공할 때까지 이동 거리를 반으로 줄이면서 제1 적대적 예제를 이동할 수 있는 것이다. 즉, 근사 방향 이동 모듈(170)은 방향 벡터 크기 등비를 축소시키면서 제1 적대적 예제를 이동시킬 수 있는 것이다.

[0073] 도 7은 일 실시예에 따른 의사결정 기반 적대적 예제 생성 방법을 설명하기 위한 흐름도이다. 도 7에 도시된 방법은 예를 들어, 전술한 적대적 예제 생성 시스템(100)에 의해 수행될 수 있다. 도시된 흐름도에서는 상기 방법을 복수 개의 단계로 나누어 기재하였으나, 적어도 일부의 단계들은 순서를 바꾸어 수행되거나, 다른 단계와 결

합되어 함께 수행되거나, 생략되거나, 세부 단계들로 나뉘어 수행되거나, 또는 도시되지 않은 하나 이상의 단계가 부가되어 수행될 수 있다.

[0074] 이하에서는, 도 7의 일부를 상세하게 설명하기 위한 흐름도인 도 8 내지 도 11을 참조하여 설명하기로 한다.

[0075] 210 단계에서, 적대적 예제 생성 시스템(100)은 제1 클래스로 분류되는 초기 적대적 예제와 제2 클래스로 분류되는 원본 데이터를 이용한 이진 탐색을 통해 제1 클래스와 제2 클래스에 대한 딥 러닝 모델의 결정 경계 상의 제1 적대적 예제를 획득할 수 있다.

[0076] 구체적으로, 적대적 예제 생성 시스템(100)은 공격 성공 판단 모듈(150)의 질의 결과를 기초로 적대적 예제와 원본 데이터 사이의 결정 경계를 이진 탐색을 이용하여 찾고, 초기 적대적 예제를 결정 경계 상의 제1 적대적 예제로 업데이트할 수 있다. 이때, 제1 적대적 예제는 초기 적대적 예제를 이동시키면서 최종 적대적 예제를 획득하는 과정에서 발생하는 결정 경계 상의 적대적 예제를 의미하는 것으로서, 초기 적대적 예제 및 최종 적대적 예제와 구분하기 위해 명명한 것일 수 있다.

[0077] 도 8을 참고하면, 211 단계에서, 적대적 예제 생성 시스템(100)은 초기 적대적 예제와 원본 데이터를 수신할 수 있다. 이때, 공격 목표 클래스로 분류되는 데이터를 초기 적대적 예제로 설정할 수 있다. 이때, 공격 목표 클래스는 타겟 클래스로도 명명될 수 있다.

[0078] 213 단계에서, 적대적 예제 생성 시스템(100)은 초기 적대적 예제가 결정 경계 상에 위치하는지 여부를 확인할 수 있다.

[0079] 213 단계의 확인 결과, 초기 적대적 예제가 결정 경계 상에 미 위치하는 경우, 215 단계에서, 적대적 예제 생성 시스템(100)은 초기 적대적 예제와 원본 데이터 간의 이진 탐색을 수행할 수 있다. 즉, 적대적 예제 생성 시스템(100)은 이진 탐색을 통해 결정 경계를 탐색하는 것이다.

[0080] 적대적 예제 생성 시스템(100)은 초기 적대적 예제가 결정 경계 상에 위치하는지 여부를 확인하는 단계 및 이진 탐색을 수행하는 단계를 반복 수행하여 결정 경계 상의 제1 적대적 예제를 획득할 수 있다. 즉, 도 2에서 도시하는 바와 같이, 적대적 예제 생성 시스템(100)은 이진 탐색을 통해 초기 적대적 예제를 결정 경계 상으로 이동시켜 제1 적대적 예제(x_t)를 획득할 수 있다. 이때, 도 2에서 도시하는 바와 같이, 설명의 편의를 위해 적대적 예제를 초기 공격 목표 클래스로 분류된 초기 적대적 예제와 결정 경계 상으로 이동시킨 제1 적대적 예제(x_t)로 구분하여 개시하기로 한다.

[0081] 220 단계에서, 적대적 예제 생성 시스템(100)은 딥 러닝 모델에 대한 적대적 공격 성공 여부에 기초하여 제1 적대적 예제와 원본 데이터 간에 직교하는 영역 내에서 제1 적대적 예제에 대한 이동 방향을 결정할 수 있다.

[0082] 도 9의 221 단계에서, 적대적 예제 생성 시스템(100)은 직교하는 영역 내 복수의 방향에 대응하는 복수의 방향 벡터를 생성할 수 있다.

[0083] 도 3과 같이, 적대적 예제 생성 시스템(100)은 $x_t - x_{orig}$ 와 직교하는 범위(A) 내에서 방향 벡터를 선택하므로 전체 공간에서 방향 벡터를 추출하는 일반적인 방식에 비해 질의 수를 줄일 수 있는 것이다. 이때, 전체 공간은 x_t 를 기준으로 전 방향을 의미하는 것일 수 있다.

[0084] 구체적으로, 적대적 예제 생성 시스템(100)은 원본 데이터로부터 시작하여 제1 적대적 예제에서 종료하는 벡터와 직교하는 영역(A) 내에서 적어도 하나 이상의 방향 벡터(V)를 산출할 수 있다.

[0085] 적대적 예제 생성 시스템(100)은 상술한 수학식 1을 기초로 복수의 방향 벡터를 생성할 수 있다. 이때, 적대적 예제 생성 시스템(100)은 방향 벡터 u 를 선택한 다음 수학식 1과 같이 $x_t - x_{orig}$ 와 수직이 되도록 처리할 수 있다.

[0086] 본 실시예는 의사결정 기반 공격기법의 기울기 예측 시에 이용하는 방향 벡터의 탐색 공간(도 3의 A)을 줄여 HopSkipJumpAttack 알고리즘에 비해 적은 질의로 방향 벡터의 예측을 수행할 수 있다는 효과를 기대할 수 있다.

[0087] 223 단계에서, 적대적 예제 생성 시스템(100)은 복수의 방향 벡터 각각과 제1 적대적 예제에 기초하여 복수의

방향 각각에 대한 적대적 공격 성공 여부를 판단할 수 있다.

[0088] 225 단계에서, 적대적 예제 생성 시스템(100)은 적대적 공격 성공 여부에 기초하여 복수의 방향 벡터 각각에 대한 가중치를 결정할 수 있다.

[0089] 적대적 예제 생성 시스템(100)은 상술한 수학식 2를 이용하여 제1 방향 벡터에 가중치를 적용할 수 있다. 이때, 적대적 예제 생성 시스템(100)은 제1 적대적 예제(x_t)가 결정 경계 상에 위치한다는 전제하게 수학식 2와 같이 몬테카를로 알고리즘을 이용하여 A_{x^*}를 근사할 수 있다.

[0090] 227 단계에서, 적대적 예제 생성 시스템(100)은 가중치 및 복수의 방향 벡터에 기초하여 이동 방향을 결정할 수 있다. 이때, 적대적 예제 생성 시스템(100)은 가중치에 기초한 복수의 방향 벡터에 대한 가중 평균을 이용하여 이동 방향에 대응하는 방향 벡터를 결정할 수 있다.

[0091] 적대적 예제 생성 시스템(100)은 가중치를 결정하는 단계에서, 복수의 방향 중 적대적 공격이 성공한 방향에 대응하는 방향 벡터에 대한 가중치를 1로 결정하고, 복수의 방향 중 적대적 공격이 실패한 방향에 대응하는 방향 벡터에 대한 가중치를 -1로 결정할 수 있다.

[0092] 230 단계에서, 적대적 예제 생성 시스템(100)은 상기 이동 방향으로 제1 적대적 예제를 이동시켜 제2 적대적 예제를 생성할 수 있다.

[0093] 적대적 예제 생성 시스템(100)은 상기 이동 방향으로 제1 적대적 예제를 기 설정된 크기만큼 이동시키되, 딥 러닝 모델에 대한 적대적 공격이 성공할 때까지 이동 거리 변경 및 이동을 반복 수행하여 제2 적대적 예제를 결정할 수 있다. 이때, 적대적 예제 생성 시스템(200)은 제1 적대적 예제를 너무 많이 이동하여 공격이 실패하는 경우 공격이 성공할 때까지 이동 거리를 반으로 줄이면서 제1 적대적 예제를 이동할 수 있는 것이다.

[0094] 도 10을 참고하면, 231 단계에서, 적대적 예제 생성 시스템(100)은 제1 적대적 예제를 기 설정된 크기만큼 이동시킬 수 있다.

[0095] 233 단계에서, 적대적 예제 생성 시스템(100)은 231 단계의 이동에 따른 적대적 공격이 성공했는지 여부를 확인할 수 있다.

[0096] 233 단계의 확인 결과 공격이 실패한 경우, 235 단계에서, 적대적 예제 생성 시스템(100)은 이동 거리를 절반으로 줄여 제1 적대적 예제를 이동시킬 수 있다. 즉, 적대적 예제 생성 시스템(100)은 방향 벡터 크기 등비를 축소시키면서 제1 적대적 예제를 이동시킬 수 있는 것이다.

[0097] 적대적 예제 생성 시스템(100)은 수학식 2와 같이, x_t가 결정 경계 상에 있다는 전제하에 $\phi_{x^*}(x_t + \delta u_b)$의 결과에 공격이 성공하는 방향은 +1, 실패하는 방향은 -1을 곱해 더한 값의 평균을 통해 대략적인 방향을 산출한 후, 수학식 3과 같이 제1 적대적 예제를 결정 경계 위로 옮기는 과정을 반복할 수 있는 것이다.

[0098] (수학식 3)

$$x_{t+1} = \alpha_t x^* + (1-\alpha_t)\left[x_t + \xi_t \frac{A_{x^*}(x_t)}{\| A_{x^*(x_t)} \|_2}\right]$$

[0099]

[0100] 이때, α_t는 라인 탐색 매개변수, x^*는 원본 데이터, ξ_t는 양의 단계 크기(positive step size), $A_{x^*}(x_t)$는 x_t에서의 공격 성공 방향일 수 있다. 여기에서, $\alpha_t \in [0,1]$로 x^*와 $\left[x_t + \xi_t \frac{A_{x^*}(x_t)}{\| A_{x^*(x_t)} \|_2}\right]$를 양 끝점으로 하는 직선 범위 내의 데이터를 나타낸다.

[0101] 구체적으로, 수학식 3을 반복하여 제1 적대적 예제를 공격 목표 클래스로 분류한 경우 1, 그렇지 않은 경우 -1을 출력하는 함수의 최적화 문제를 해결할 수 있다. 현재 반복에서의 제1 적대적 예제 x_t에서의 공격 성공 방향

$A_{x^*}(x_t)$를 산출하는 것은 도 3과 같을 수 있고, $A_{x^*}(x_t)$의 방향에 ξ_t만큼 곱하여 x_t에 더하는 것은 도 4와 같을 수 있다. 이때, ξ_t는 제1 적대적 예제의 분류가 변하지 않도록 조절하는 것으로서, 분류가 변하는 경우 반씩 나눠 크기를 줄일 수 있다. 다음, 도 2와 같이 제1 적대적 예제를 결정 경계 상으로 이동시키는데, 원본 데이터 x^*와 $\left\{x_t+\xi_t\frac{A_{x^*}(x_t)}{\| A_{x^*(x_t)} \|_2}\right\}$ 사이의 결정 경계에 위치하는 다음 차수의 제1 적대적 예제인 x_{t+1}을 산출할 수 있다. 이때, 이진 탐색을 이용하여 적절한 값의 α_t를 획득할 수 있다. 수학식 3의 경우 x_t가 결정 경계 위에 있다는 전제하에 수학식 2와 같이 몬테카를로 알고리즘을 이용하여 A_{x^*}를 근사할 수 있다.

[0102] 본 실시예의 알고리즘은 도 6과 같을 수 있다. 실험 과정 중 각 반복에서 기울기를 예측하는데 이용하는 질의를 줄이는 만큼 전체 반복 수를 증가시켜 실험을 진행하였다. HopSkipJumpAttack의 경우 반복 수가 늘어날수록 보다 정확한 기울기 예측을 위해 더 많은 방향 벡터를 이용하는데 본 실시예에서는 반복마다 HopSkipJumpAttack에 비해 적은 수(예를 들어, 100개)의 방향 벡터를 적용할 수 있었다.

[0103] 240 단계에서, 적대적 예제 생성 시스템(100)은 원본 데이터에 인접한 결정 경계 상에 위치하는 최종 적대적 예제를 결정할 수 있다.

[0104] 도 11을 참고하면, 241 단계에서, 적대적 예제 생성 시스템(100)은 제1 적대적 예제가 원본 데이터에 인접한 결정 경계 상에 위치하는지 여부를 확인할 수 있다. 이때, 원본 데이터에 인접한 결정 경계는 사전에 임의로 설정될 수 있다.

[0105] 241 단계의 확인 결과 결정 경계 상에 위치하는 경우, 243 단계에서, 적대적 예제 생성 시스템(100)은 해당 제1 적대적 예제를 최종 적대적 예제로 결정할 수 있다.

[0106] 한편, 241 단계의 확인 결과 결정 경계 상에 미 위치하는 경우, 245 단계에서, 적대적 예제 생성 시스템(100)은 제1 적대적 예제와 원본 데이터 간의 이진 탐색을 수행하여 원본 데이터에 인접한 결정 경계 상에 위치하는 최종 적대적 예제를 획득할 수 있다.

[0107] 단계 240에서, 적대적 예제 생성 시스템(100)은 원본 데이터에 인접한 결정 경계 상에 위치하되, 공격 목표 클래스 분류를 유지하고 원본 데이터와의 차이가 기 설정된 기준 이하로 작은 최종 적대적 예제를 산출할 수 있다.

[0108] 250 단계에서, 적대적 예제 생성 시스템(100)은 최종 적대적 예제를 이용하여 딥 러닝 모델로 공격을 수행할 수 있다.

[0109] 한편, 적대적 예제 생성 시스템(100)은 기정된 종료 조건을 만족할 때까지 상기 제2 적대적 예제를 상기 초기 적대적 예제로 이용하여 상기 제1 적대적 예제를 획득하는 단계 210 내지 상기 제2 적대적 예제를 생성하는 단계 230를 반복 수행할 수 있다.

[0111] 도 12는 일 실시예에 따른 컴퓨팅 장치를 포함하는 컴퓨팅 환경을 예시하여 설명하기 위한 블록도이다. 도시된 실시예에 서, 각 컴포넌트들은 이하에 기술된 것 이외에 상이한 기능 및 능력을 가질 수 있고, 이하에 기술된 것 이외에도 추가적인 컴포넌트를 포함할 수 있다.

[0112] 도시된 컴퓨팅 환경(10)은 컴퓨팅 장치(12)를 포함한다. 일 실시예에서, 컴퓨팅 장치(12)는 적대적 예제 생성 시스템(100)일 수 있다.

[0113] 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 동작

들을 수행하도록 구성될 수 있다.

[0114] 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.

[0115] 통신 버스(18)는 프로세서(14), 컴퓨터 판독 가능 저장 매체(16)를 포함하여 컴퓨팅 장치(12)의 다른 다양한 컴포넌트들을 상호 연결한다.

[0116] 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.

[0118] 상술한 본 실시예는 의사결정 기반 공격 기법의 많은 질의 수가 필요한 문제점을 개선하여 상용화된 딥 러닝 모델에 대해 공격을 수행할 수 있는 적대적 예제 생성 시스템을 개시한다. 딥 러닝 모델은 현재 음성 인식, 자율주행차 등 우리 생활에 밀접하게 관련되어 보안 문제 발생 시 민감한 정보 누출, 인명피해 등을 초래할 수 있다. 따라서 딥 러닝 모델의 잘 알려진 취약점인 적대적 예제를 통해 더욱 강건한 딥 러닝 모델 설계 및 안전성 검토가 필요하다.

[0119] 본 실시예에서 제안한 적대적 예제 생성 시스템은 방향 벡터의 탐색 공간을 줄이는 방법을 사용한다. HopSkipJump 공격의 경우 기울기를 근사할 때 몬테카를로 알고리즘을 이용하는데 이는 결정경계 상에 있는 적대적 예제에 가능한 모든 데이터 범위의 랜덤 방향벡터를 더하여 모델에 질의한 다음 질의 결과를 바탕으로 기울기를 근사하는 방식이다. 이때, 수백 개의 질의가 소모되는데 이미지, 오디오 데이터는 크기가 매우 크기 때문에 질의를 무한히 하지 않는 한 정확한 기울기 근사는 어려울 수 있다. 본 실시예는 정확한 기울기 근사 대신 랜덤 방향벡터의 탐색 공간을 줄여 대략적인 공격 성공 방향을 구하는 방식을 적용하기 때문에, 질의 수를 줄일 수 있다는 효과를 기대할 수 있다. 구체적으로, 방향 벡터의 탐색 공간 축소를 위해 원본 데이터와 결정 경계 상의 적대적 예제와 직교하는 범위 내의 공간에서만 방향 벡터를 추출하여 질의 낭비를 최소화 하는 것이다. 각 반복 단계에서 기울기를 예측하는데 필요한 질의 수를 줄여 공격이 성공하는 대략적인 방향을 구하는 것만으로도 적대적 예제 생성이 가능하고 절약한 질의 수 만큼 최적화 반복 수를 늘리면 같은 질의 수 대비 HopSkipJump 공격보다 원본데이터와 더 유사한 적대적 예제를 생성할 수 있는 것이다. 즉, 본 실시예는 의사 결정기반 공격 기법의 기울기 예측 시 이용되는 방향 벡터의 탐색 공간 축소 방안을 적용하여 질의 효율성을 증대시키므로 실제 환경에서 적용될 수 있는 적대적 공격을 수행할 수 있는 것이다.

[0120] 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 청구범위뿐만 아니라 이 청구범위와 균등한 것들에 의해 정해져야 한다.

부호의 설명

[0121] 10: 컴퓨팅 환경

12: 컴퓨팅 장치

14: 프로세서

16: 컴퓨터 판독 가능 저장 매체

18: 통신 버스

20: 프로그램

22: 입출력 인터페이스

24: 입출력 장치

26: 네트워크 통신 인터페이스

100: 적대적 예제 생성 시스템

110: 결정 경계 이동 모듈

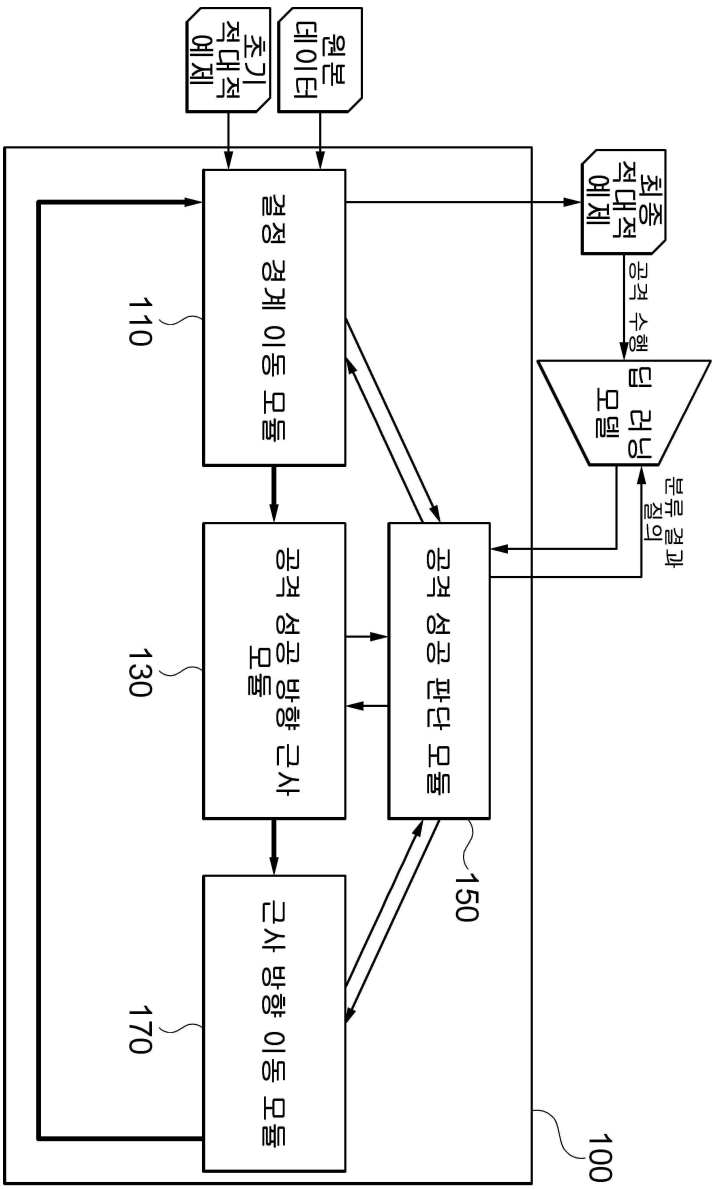
130: 공격 성공 방향 근사 모듈

150: 공격 성공 판단 모듈

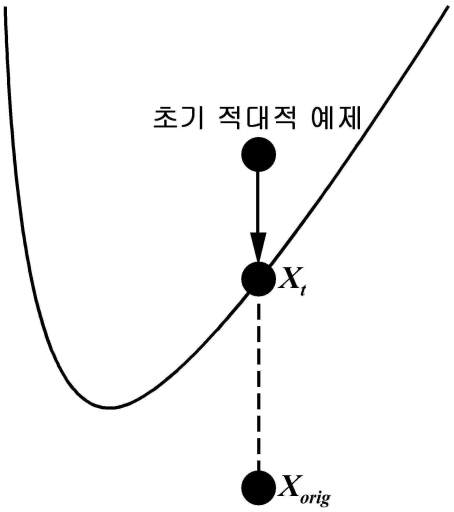
170: 근사 방향 이동 모듈

도면

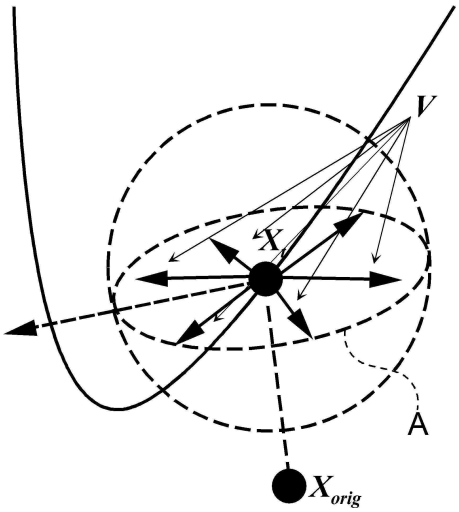
도면1


초기 적대적 예제
원본 데이터
결정 경계 이동 모듈
110
최종 적대적 예제
공격 수행
딥 러닝 모델
분류 결과 질의
공격 성공 방향 근사 모듈
130
공격 성공 판단 모듈
150
근사 방향 이동 모듈
170
100

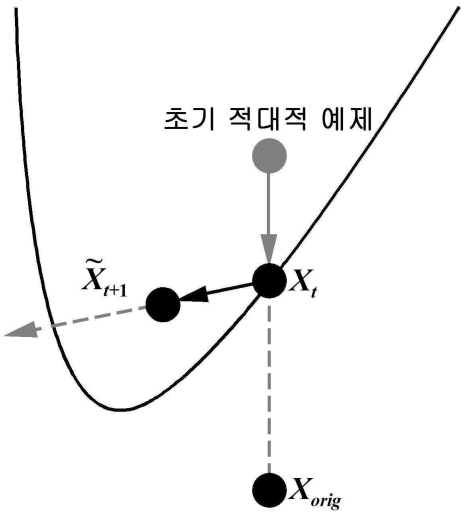
도면2



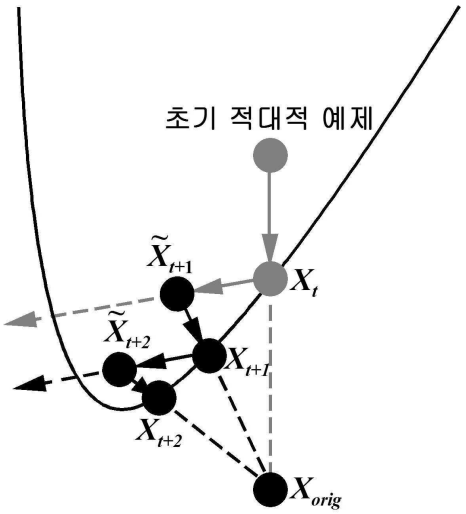
도면3



도면4



도면5



도면6

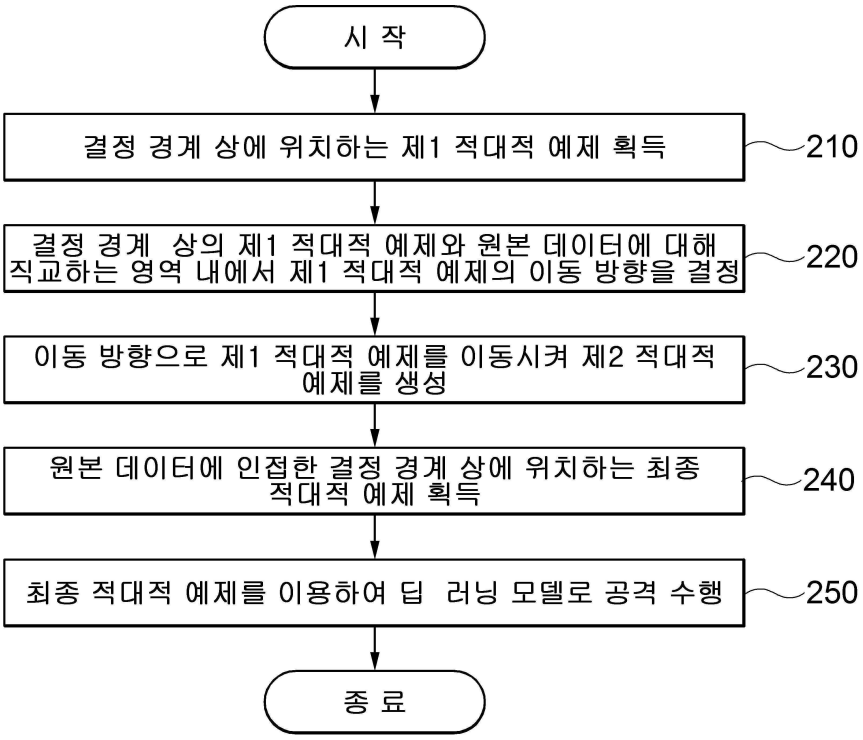
Algorithm Generating audio AEs.

Inputs: Classifier C. Original audio x^*. Initial adversarial example $\tilde{x}_0$. Iterations T

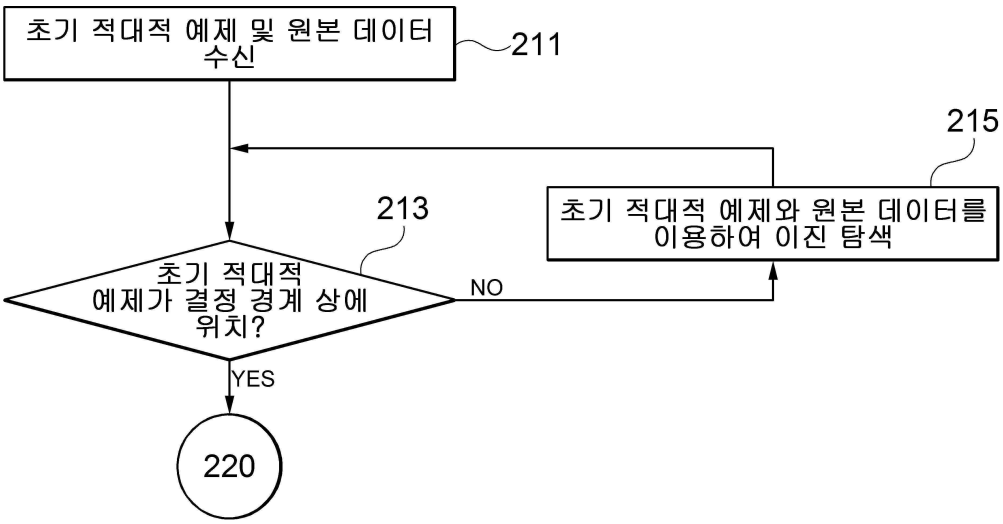
Outputs: The adversarial example x_t

```
for t in 1,2,....,T-1 do
    # Boundary search
    x_t = Bin-search(x̃_{t-1}, x*)
    # Reducing the search space
    for i in 100 do
        Set random vector u_i
        u_i = u_i − proj_{x_t − x*} u_i
    end for
    # Gradient estimation
    ∇̃S = Grad-estimation(x*, x_t, U)
    x̃_t = x_t + ∇̃S / ‖∇̃S‖
end for
return x_t = Bin-search(x̃_{t-1}, x*)
```

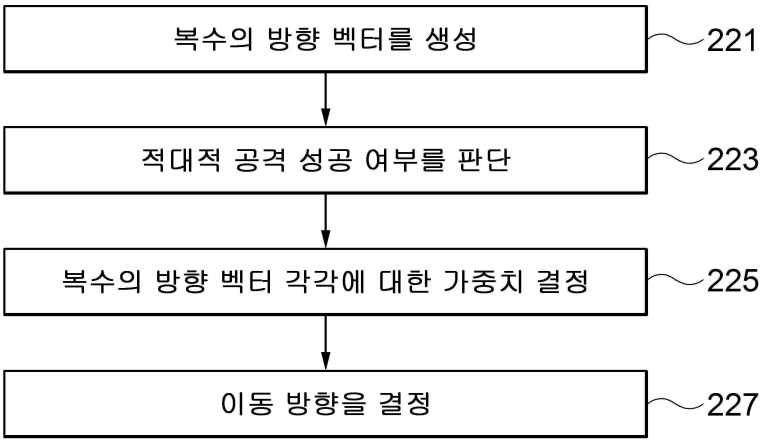
도면7


시 작
결정 경계 상에 위치하는 제1 적대적 예제 획득
210
결정 경계 상의 제1 적대적 예제와 원본 데이터에 대해 직교하는 영역 내에서 제1 적대적 예제의 이동 방향을 결정
220
이동 방향으로 제1 적대적 예제를 이동시켜 제2 적대적 예제를 생성
230
원본 데이터에 인접한 결정 경계 상에 위치하는 최종 적대적 예제 획득
240
최종 적대적 예제를 이용하여 딥 러닝 모델로 공격 수행
250
종 료

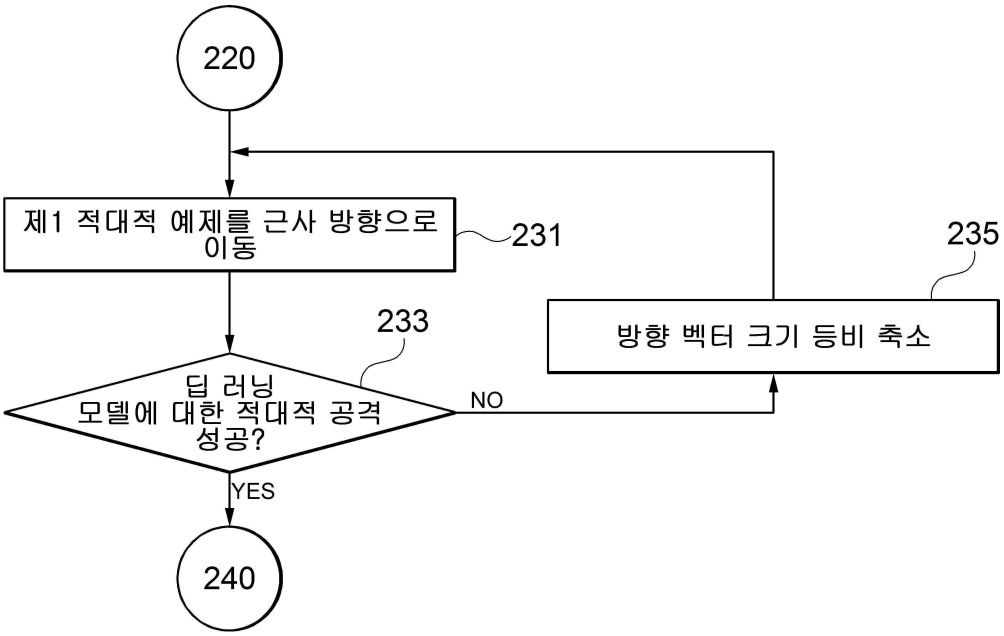
도면8


초기 적대적 예제 및 원본 데이터 수신
211
215
213
초기 적대적 예제와 원본 데이터를 이용하여 이진 탐색
초기 적대적 예제가 결정 경계 상에 위치?
NO
YES
220

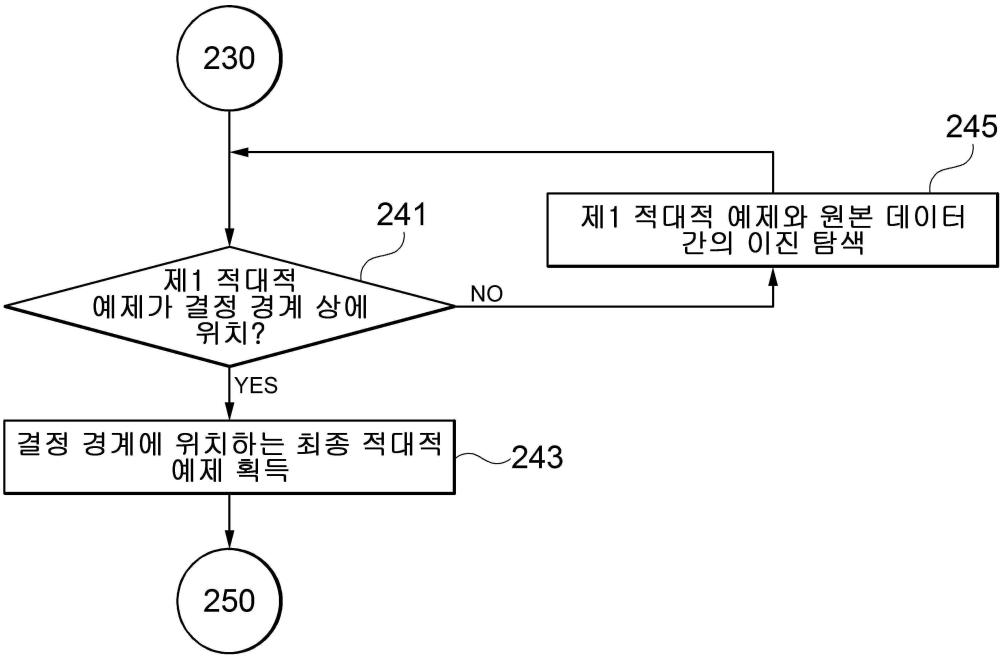
도면9


복수의 방향 벡터를 생성
221
적대적 공격 성공 여부를 판단
223
복수의 방향 벡터 각각에 대한 가중치 결정
225
이동 방향을 결정
227

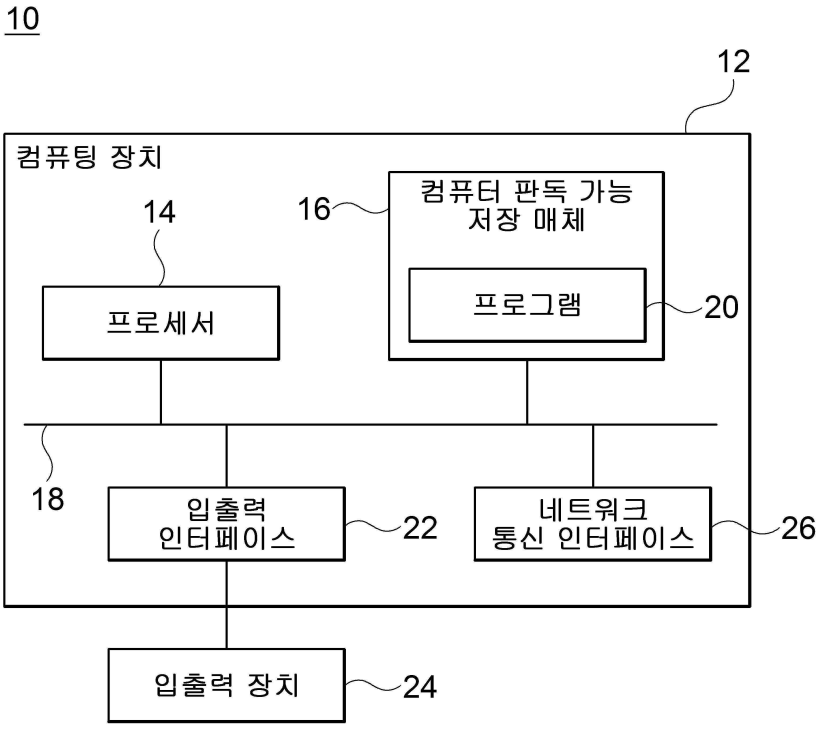
도면10


220
제1 적대적 예제를 근사 방향으로 이동
231
235
방향 벡터 크기 등비 축소
233
딥 러닝 모델에 대한 적대적 공격 성공?
NO
YES
240

도면11


230
245
241
제1 적대적 예제와 원본 데이터 간의 이진 탐색
제1 적대적 예제가 결정 경계 상에 위치?
NO
YES
결정 경계에 위치하는 최종 적대적 예제 획득
243
250

도면12


10
12
컴퓨팅 장치
14
16
컴퓨터 판독 가능 저장 매체
프로세서
프로그램
20
18
입출력 인터페이스
22
네트워크 통신 인터페이스
26
입출력 장치
24