



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0018725
(43) 공개일자 2025년02월07일

(51) 국제특허분류(Int. Cl.)
G06N 3/094 (2023.01) G06N 3/0475 (2023.01)
(52) CPC특허분류
G06N 3/094 (2023.01)
G06N 3/0475 (2023.01)
(21) 출원번호 10-2023-0099706
(22) 출원일자 2023년07월31일
심사청구일자 2023년07월31일

(71) 출원인
경희대학교 산학협력단
경기도 용인시 기흥구 덕영대로 1732 (서천동, 경희대학교 국제캠퍼스내)
(72) 발명자
황효석
서울특별시 서초구 잠원로 85 (잠원동, 블루힐하우스)
유근혁
경기도 수원시 영통구 매영로 433, 205호 (영통동)
전민우
경기도 수원시 영통구 봉영로1482번길 7-13, 102호 (영통동)
(74) 대리인
두호특허법인

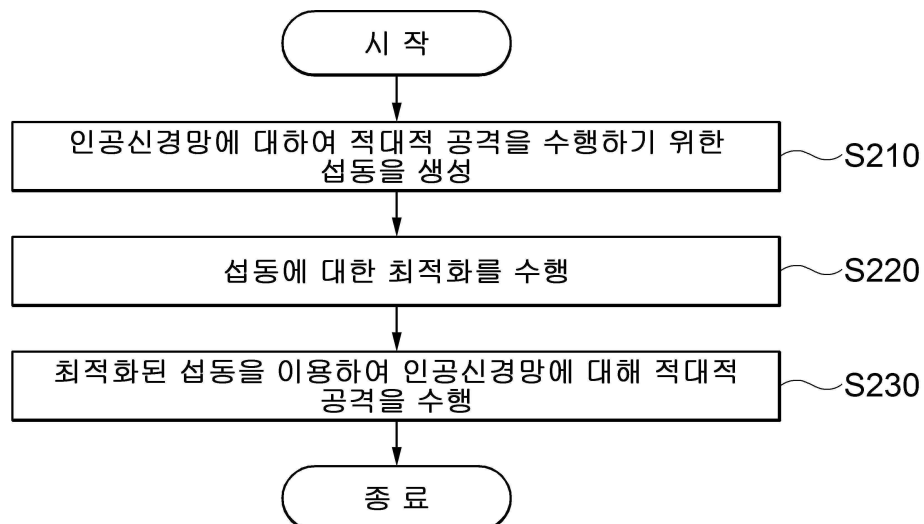
전체 청구항 수 : 총 14 항

(54) 발명의 명칭 **인공 신경망 모델의 적대적 공격 방법 및 이를 수행하기 위한 컴퓨팅 장치**

(57) 요약

인공 신경망 모델의 적대적 공격 방법 및 이를 수행하기 위한 컴퓨팅 장치가 개시된다. 본 발명의 일 실시예에 따른 인공 신경망 모델의 적대적 공격 방법 및 이를 수행하기 위한 컴퓨팅 장치는 하나 이상의 프로세서들, 및 상기 하나 이상의 프로세서들에 의해 실행되는 하나 이상의 프로그램들을 저장하는 메모리를 구비한 컴퓨팅 장치에서 수행되는 인공 신경망 모델의 적대적 공격 방법으로서, 인공신경망에 대하여 적대적 공격을 수행하기 위한 섭동을 생성하는 단계; 상기 생성된 섭동에 대한 최적화를 수행하는 단계; 및 상기 최적화된 섭동을 이용하여 상기 인공신경망에 대한 적대적 공격을 수행하는 단계를 포함한다.

대표도 - 도2



이 발명을 지원한 국가연구개발사업

과제고유번호	1711179355
과제번호	00167169
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신방송기술국제공동연구
연구과제명	이동형 로봇 기반 실사 메타버스 실감형 비디오의 획득및 처리 기술 개발
기 여 율	1/1
과제수행기관명	성균관대학교산학협력단
연구기간	2023.01.01 ~ 2023.12.31

명세서

청구범위

청구항 1

하나 이상의 프로세서들, 및

상기 하나 이상의 프로세서들에 의해 실행되는 하나 이상의 프로그램들을 저장하는 메모리를 구비한 컴퓨팅 장치에서 수행되는 인공 신경망 모델의 적대적 공격 방법으로서,

인공신경망에 대하여 적대적 공격(Adversarial Attack)을 수행하기 위한 섭동(perturbation)을 생성하는 단계;

상기 생성된 섭동에 대한 최적화를 수행하는 단계; 및

상기 최적화된 섭동을 이용하여 상기 인공신경망에 대한 적대적 공격을 수행하는 단계를 포함하는, 인공 신경망 모델의 적대적 공격 방법.

청구항 2

청구항 1에 있어서,

상기 섭동을 생성하는 단계는,

원본 데이터 세트에 포함된 이미지와 동일한 크기로 상기 섭동을 생성하는 단계; 및

상기 생성된 섭동을 초기화하는 단계를 더 포함하는, 인공 신경망 모델의 적대적 공격 방법.

청구항 3

청구항 1에 있어서,

상기 최적화를 수행하는 단계는,

원본 데이터 세트에 상기 섭동을 추가하여 제1 적대적 데이터 세트를 생성하는 단계;

상기 원본 데이터 세트에 상기 섭동 및 노이즈를 추가하여 제2 적대적 데이터 세트를 생성하는 단계;

상기 제1 적대적 데이터 세트 및 상기 제2 적대적 데이터 세트를 상기 인공 신경망에 입력하는 단계;

상기 인공 신경망의 출력 값을 기반으로 손실함수를 이용하여 상기 제1 적대적 데이터 세트에 대한 제1 성능 지표 및 상기 제2 적대적 데이터 세트에 대한 제2 성능 지표를 각각 산출하는 단계;

상기 제1 성능 지표 및 상기 제2 성능 지표를 기반으로 RGF(Random Gradient Free) 알고리즘을 이용하여 섭동에 대한 기울기를 산출하는 단계; 및

상기 섭동에 대한 기울기 및 상기 노이즈를 상기 섭동에 반영하여 상기 섭동을 업데이트하는 단계를 더 포함하는, 인공 신경망 모델의 적대적 공격 방법.

청구항 4

청구항 3에 있어서,

상기 수학식 2에 의해 성능 지표를 산출하는, 인공 신경망 모델의 적대적 공격 방법.

[수학식 2]

$$L_H(\mathbf{y}_0, \mathbf{y}_1) = \frac{1}{N_{cls}} \sum_{i=1}^{N_{cls}} \max(0, y_{0,i} + y_{1,i} - 1)$$

$$L_H(\mathbf{y}_u, \mathbf{y}_1) = \frac{1}{N_{cls}} \sum_{i=1}^{N_{cls}} \max(0, y_{u,i} + y_{1,i} - 1)$$

(여기서, $L_H(\mathbf{y}_0, \mathbf{y}_1)$ 는 제1 성능 지표, $L_H(\mathbf{y}_u, \mathbf{y}_1)$ 는 제2 성능 지표, $\mathbf{y}_0$ 는 제1 적대적 데이터 세트에 대한 인공 신경망의 출력 값, $\mathbf{y}_u$ 는 제2 적대적 데이터 세트에 대한 인공 신경망의 출력 값, $\mathbf{y}_1$ 는 원본 데이터 세트에 대한 인공 신경망의 출력 값, N_{cls} 는 출력될 수 있는 클래스(출력 값)의 개수, i 는 인공 신경망에서 출력되는 클래스의 각 번호)

청구항 5

청구항 4에 있어서,

하기 수학식 3에 의해 상기 섭동에 대한 기울기를 산출하는, 인공 신경망 모델의 적대적 공격 방법.

[수학식 3]

$$g = \frac{L_H(\hat{\mathbf{y}}_u, \mathbf{y}) - L_H(\hat{\mathbf{y}}_0, \mathbf{y})}{\gamma}$$

(여기서, g 는 섭동에 대한 기울기, $L_H(\hat{\mathbf{y}}_0, \mathbf{y})$ 는 제1 성능 지표, $L_H(\hat{\mathbf{y}}_u, \mathbf{y})$ 는 제2 성능 지표, γ 은 기울기 계산을 위한 상수, $\hat{\mathbf{y}}_u$ 는 제2 적대적 데이터 세트에 대한 인공 신경망의 출력 값, $\hat{\mathbf{y}}_0$ 는 제1 적대적 데이터 세트에 대한 인공 신경망의 출력 값, $\mathbf{y}$ 는 원본 데이터 세트에 대한 인공 신경망의 출력 값)

청구항 6

청구항 5에 있어서,

하기 수학식 4에 의해 상기 섭동을 업데이트하는, 인공 신경망 모델의 적대적 공격 방법.

[수학식 4]

$$\mathbf{p}_{t+1} = \mathbf{p}_t + \alpha \cdot \mathbf{u}_t \cdot \frac{L_H(\hat{\mathbf{y}}_{u_t}, \mathbf{y}) - L_H(\hat{\mathbf{y}}_0, \mathbf{y})}{\gamma}$$

(여기서, $\mathbf{p}_{t+1}$ 는 업데이트된 섭동, $\mathbf{p}_t$ 는 t 번째 섭동, α 는 업데이트 단계에 따른 크기 값, $\mathbf{u}_t$ 는 t 번째 노이즈)

청구항 7

청구항 3에 있어서,

상기 원본 데이터 세트는, 배치 사이즈(Batch size)로 구성되며,

상기 노이즈는, 가우시안 노이즈(Gaussian Noise)인, 인공 신경망 모델의 적대적 공격 방법.

청구항 8

하나 이상의 프로세서들;

메모리; 및

하나 이상의 프로그램들을 포함하고,

상기 하나 이상의 프로그램들은 상기 메모리에 저장되고, 상기 하나 이상의 프로세서들에 의해 실행되도록 구성되며,

상기 하나 이상의 프로그램들은,

인공신경망에 대하여 적대적 공격(Adversarial Attack)을 수행하기 위한 섭동(perturbation)을 생성하기 위한 명령;

상기 생성된 섭동에 대한 최적화를 수행하기 위한 명령; 및

상기 최적화된 섭동을 이용하여 상기 인공신경망에 대한 적대적 공격을 수행하기 위한 명령을 포함하는, 컴퓨팅 장치.

청구항 9

청구항 8에 있어서,

상기 섭동을 생성하기 위한 명령은,

원본 데이터 세트에 포함된 이미지와 동일한 크기로 상기 섭동을 생성하기 위한 명령; 및

상기 생성된 섭동을 초기화하기 위한 명령을 더 포함하는, 컴퓨팅 장치.

청구항 10

청구항 8에 있어서,

상기 최적화를 수행하기 위한 명령은,

원본 데이터 세트에 상기 섭동을 추가하여 제1 적대적 데이터 세트를 생성하기 위한 명령;

상기 원본 데이터 세트에 상기 섭동 및 노이즈를 추가하여 제2 적대적 데이터 세트를 생성하기 위한 명령;

상기 제1 적대적 데이터 세트 및 상기 제2 적대적 데이터 세트를 상기 인공 신경망에 입력하기 위한 명령;

상기 인공 신경망의 출력 값을 기반으로 손실함수를 이용하여 상기 제1 적대적 데이터 세트에 대한 제1 성능 지표 및 상기 제2 적대적 데이터 세트에 대한 제2 성능 지표를 각각 산출하기 위한 명령;

상기 제1 성능 지표 및 상기 제2 성능 지표를 기반으로 RGF(Random Gradient Free) 알고리즘을 이용하여 섭동에 대한 기울기를 산출하기 위한 명령; 및

상기 섭동에 대한 기울기 및 상기 노이즈를 상기 섭동에 반영하여 상기 섭동을 업데이트하기 위한 명령을 더 포함하는, 컴퓨팅 장치..

청구항 11

청구항 10에 있어서,

상기 수학식 2에 의해 성능 지표를 산출하는, 컴퓨팅 장치.

[수학식 2]

$$L_H(\mathbf{y}_0, \mathbf{y}_1) = \frac{1}{N_{cls}} \sum_{i=1}^{N_{cls}} \max(0, y_{0,i} + y_{1,i} - 1)$$

$$L_H(\mathbf{y}_u, \mathbf{y}_1) = \frac{1}{N_{cls}} \sum_{i=1}^{N_{cls}} \max(0, y_{u,i} + y_{1,i} - 1)$$

(여기서, $L_H(\mathbf{y}_0, \mathbf{y}_1)$ 는 제1 성능 지표, $L_H(\mathbf{y}_u, \mathbf{y}_1)$ 는 제2 성능 지표, $\mathbf{y}_0$ 는 제1 적대적 데이터 세트에 대한 인공 신경망의 출력 값, $\mathbf{y}_u$ 는 제2 적대적 데이터 세트에 대한 인공 신경망의 출력 값, $\mathbf{y}_1$ 는 원본 데이터 세트에 대한 인공 신경망의 출력 값, N_{cls} 는 출력될 수 있는 클래스(출력 값)의 개수, i 는 인공 신경망에서 출력되는 클래스의 각 번호)

청구항 12

청구항 11에 있어서,

하기 수학식 3에 의해 상기 섭동에 대한 기울기를 산출하는, 컴퓨팅 장치.

[수학식 3]

$$g = \frac{L_H(\hat{\mathbf{y}}_u, \mathbf{y}) - L_H(\hat{\mathbf{y}}_0, \mathbf{y})}{\gamma}$$

(여기서, g 는 섭동에 대한 기울기, $L_H(\hat{\mathbf{y}}_u, \mathbf{y})$ 는 제1 성능 지표, $L_H(\hat{\mathbf{y}}_0, \mathbf{y})$ 는 제2 성능 지표, γ 은 기울기 계산을 위한 상수, $\hat{\mathbf{y}}_u$ 는 제2 적대적 데이터 세트에 대한 인공 신경망의 출력 값, $\hat{\mathbf{y}}_0$ 는 제1 적대적 데이터 세트에 대한 인공 신경망의 출력 값, $\mathbf{y}$ 는 원본 데이터 세트에 대한 인공 신경망의 출력 값)

청구항 13

청구항 12에 있어서,

하기 수학식 4에 의해 상기 섭동을 업데이트하는, 컴퓨팅 장치.

[수학식 4]

$$\mathbf{p}_{t+1} = \mathbf{p}_t + \alpha \cdot \mathbf{u}_t \cdot \frac{L_H(\hat{\mathbf{y}}_{u_t}, \mathbf{y}) - L_H(\hat{\mathbf{y}}_0, \mathbf{y})}{\gamma}$$

(여기서, $\mathbf{p}_{t+1}$ 는 업데이트된 섭동, $\mathbf{p}_t$ 는 t 번째 섭동, α 는 업데이트 단계에 따른 크기 값, $\mathbf{u}_t$ 는 t 번째 노이즈)

청구항 14

청구항 10에 있어서,

상기 원본 데이터 세트는, 배치 사이즈(Batch size)로 구성되며,

상기 노이즈는, 가우시안 노이즈(Gaussian Noise)인, 컴퓨팅 장치.

발명의 설명

기술분야

[0001] 본 발명의 실시예들은 인공 신경망 모델의 적대적 공격 기술과 관련된다.

배경기술

[0003] 심층 신경망(Deep Neural Networks; DNN)의 은닉 레이어가 출력하는 값은 심층 신경망의 입력 값에 종속된다. 심층 신경망의 입력 값을 적절히 변경하면 은닉 레이어는 전혀 다른 특징을 추출하게 되고, 출력 레이어가 다른 의사결정(decision)을 내리게 할 수 있다. 이렇게 입력 값에 사람이 인식할 수 없는 노이즈를 추가하여 인간의 인식에는 아무런 변화가 없지만 심층 신경망이 잘못 분류(다른 의사결정)하도록 하는 공격을 적대적 공격(Adversarial Attack)이라 한다. 이 때, 원래 입력 값과 변경된 입력 값의 차이(즉, 원래 입력 값에 추가하는 노이즈)를 섭동(perturbation)이라 한다.

[0004] 적대적 공격은 크게 심층 신경망의 매개 변수, 손실 함수, 구조 등 그 내부 정보를 사용하는 화이트 박스 공격(white box attack)과 심층 신경망의 내부 정보 없이 오직 심층 신경망의 입력과 출력 값만을 가지고 수행하는 블랙 박스 공격(black box attack)으로 구분할 수 있다. 그 중에서도 심층 신경망의 내부 정보를 알 수 없는 블랙박스 적대적 공격은 더욱 실용적이고 어려운 공격으로 평가된다. 이러한 블랙박스 적대적 공격은 카테고리별 점수를 알 때는 점수 기반 공격(score-based attack)과 카테고리별 점수를 모를 때는 의사결정 기반 공격(decision-based attack)으로 구분할 수 있으며, 의사결정 기반 공격은 난이도 관점에서 가장 어려운 공격으로 평가된다.

[0005] 즉, 적대적 공격은 대상 심층 신경망에 대한 정보가 많을수록 쉽다. 첫째로 대상 심층 신경망의 자세한 구조를 안다면 어떤 섭동이 중요한 특징을 추출하지 못하게 만드는지 확인할 수 있고, 둘째로, 출력 레이어의 점수를 안다면 어떤 카테고리로 잘못 분류하도록 섭동을 최적화하는 게 효율적인지 알 수 있다.

[0006] 심층 신경망은 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 평가받아야 한다. 여기서, 적대적 공격은 심층 신경망의 강건성을 평가하고 개선하는데 매우 중요한 역할을 하게 된다. 즉, 효율적인 심층 신경망의 강건성 평가를 위해서 높은 공격 성공률과 상대적으로 적은 자원으로도 공격을 수행할 수 있는 적대적 공격 방법(즉, 섭동이 최적화된 적대적 공격 방법)이 필요하다.

선행기술문헌

특허문헌

[0008] (특허문헌 0001) 한국공개특허공보 제10-2022-0025455호(2022.03.03)

발명의 내용

해결하려는 과제

[0009] 본 발명의 실시예들은 상대적으로 적은 자원으로 공격 성공률을 높일 수 있는 적대적 공격 방법을 제공하기 위한 것이다.

과제의 해결 수단

[0011] 본 발명의 예시적인 실시예에 따르면, 하나 이상의 프로세서들, 및 상기 하나 이상의 프로세서들에 의해 실행되는 하나 이상의 프로그램들을 저장하는 메모리를 구비한 컴퓨팅 장치에서 수행되는 인공 신경망 모델의 적대적 공격 방법으로서, 인공신경망에 대하여 적대적 공격을 수행하기 위한 섭동을 생성하는 단계, 상기 생성된 섭동에 대한 최적화를 수행하는 단계 및 상기 최적화된 섭동을 이용하여 상기 인공신경망에 대한 적대적 공격을 수

행하는 단계를 포함하는 인공 신경망 모델의 적대적 공격 방법이 제공된다.

[0012] 상기 섭동을 생성하는 단계는 원본 데이터 세트에 포함된 이미지와 동일한 크기로 상기 섭동을 생성하는 단계 및 상기 생성된 섭동을 초기화하는 단계를 더 포함할 수 있다.

[0013] 상기 최적화를 수행하는 단계는 원본 데이터 세트에 상기 섭동을 추가하여 제1 적대적 데이터 세트를 생성하는 단계, 상기 원본 데이터 세트에 상기 섭동 및 노이즈를 추가하여 제2 적대적 데이터 세트를 생성하는 단계, 상기 제1 적대적 데이터 세트 및 상기 제2 적대적 데이터 세트를 상기 인공 신경망에 입력하는 단계, 상기 인공 신경망의 출력 값을 기반으로 손실함수를 이용하여 상기 제1 적대적 데이터 세트에 대한 제1 성능 지표 및 상기 제2 적대적 데이터 세트에 대한 제2 성능 지표를 각각 산출하는 단계, 상기 제1 성능 지표 및 상기 제2 성능 지표를 기반으로 RGF(Random Gradient Free) 알고리즘을 이용하여 섭동에 대한 기울기를 산출하는 단계 및 상기 섭동에 대한 기울기 및 상기 노이즈를 상기 섭동에 반영하여 상기 섭동을 업데이트하는 단계를 더 포함할 수 있다.

[0014] 하기 수학적 식 2에 의해 성능 지표를 산출할 수 있다.

[0015] [수학적 식 2]

$$L_H(\mathbf{y}_0, \mathbf{y}_1) = \frac{1}{N_{cls}} \sum_{i=1}^{N_{cls}} \max(0, y_{0,i} + y_{1,i} - 1)$$

$$L_H(\mathbf{y}_u, \mathbf{y}_1) = \frac{1}{N_{cls}} \sum_{i=1}^{N_{cls}} \max(0, y_{u,i} + y_{1,i} - 1)$$

[0018] (여기서, $L_H(\mathbf{y}_0, \mathbf{y}_1)$ 는 제1 성능 지표, $L_H(\mathbf{y}_u, \mathbf{y}_1)$ 는 제2 성능 지표, $\mathbf{y}_0$ 는 제1 적대적 데이터 세트에 대한 인공 신경망의 출력 값, $\mathbf{y}_u$ 는 제2 적대적 데이터 세트에 대한 인공 신경망의 출력 값, $\mathbf{y}_1$ 는 원본 데이터 세트에 대한 인공 신경망의 출력 값, N_{cls} 는 출력될 수 있는 클래스(출력 값)의 개수, i 는 인공 신경망에서 출력되는 클래스의 각 번호)

[0019] 하기 수학적 식 3에 의해 상기 섭동에 대한 기울기를 산출할 수 있다.

[0020] [수학적 식 3]

$$g = \frac{L_H(\hat{\mathbf{y}}_u, \mathbf{y}) - L_H(\hat{\mathbf{y}}_0, \mathbf{y})}{\gamma}$$

[0022] (여기서, g 는 섭동에 대한 기울기, $L_H(\hat{\mathbf{y}}_0, \mathbf{y})$ 는 제1 성능 지표, $L_H(\hat{\mathbf{y}}_u, \mathbf{y})$ 는 제2 성능 지표, γ 은 기울기 계산을 위한 상수, $\hat{\mathbf{y}}_u$ 는 제2 적대적 데이터 세트에 대한 인공 신경망의 출력 값, $\hat{\mathbf{y}}_0$ 는 제1 적대적 데이터 세트에 대한 인공 신경망의 출력 값, $\mathbf{y}$ 는 원본 데이터 세트에 대한 인공 신경망의 출력 값)

[0023] 하기 수학적 식 4에 의해 상기 섭동을 업데이트할 수 있다.

[0024] [수학적 식 4]

$$\mathbf{p}_{t+1} = \mathbf{p}_t + \alpha \cdot \mathbf{u}_t \cdot \frac{L_H(\hat{\mathbf{y}}_{u_t}, \mathbf{y}) - L_H(\hat{\mathbf{y}}_0, \mathbf{y})}{\gamma}$$

[0026] (여기서, $\mathbf{p}_{t+1}$ 는 업데이트된 섭동, $\mathbf{p}_t$ 는 t 번째 섭동, α 는 업데이트 단계에 따른 크기 값, $\mathbf{u}_t$ 는 t 번째 노이즈)

[0027] 상기 원본 데이터 세트는 배치 사이즈(Batch size)로 구성되며, 상기 노이즈는, 가우시안 노이즈(Gaussian Noise)일 수 있다.

[0028] 본 발명의 다른 예시적인 실시예에 따르면, 하나 이상의 프로세서들; 메모리; 및 하나 이상의 프로그램들을 포함하고, 상기 하나 이상의 프로그램들은 상기 메모리에 저장되고, 상기 하나 이상의 프로세서들에 의해 실행되도록 구성되며, 상기 하나 이상의 프로그램들은, 인공신경망에 대하여 적대적 공격을 수행하기 위한 섭동을 생성하기 위한 명령, 상기 생성된 섭동에 대한 최적화를 수행하기 위한 명령 및 상기 최적화된 섭동을 이용하여

상기 인공지능망에 대한 적대적 공격을 수행하기 위한 명령을 포함하는 컴퓨팅 장치가 제공된다.

발명의 효과

- [0030] 본 발명의 실시예들에 따르면, 원본 데이터 세트에 추가되는 섭동을 최적화함으로써, 높은 공격 성공률과 상대적으로 적은 자원으로도 적대적 공격을 수행할 수 있게 된다.
- [0031] 또한, 본 발명의 실시예들에 따르면, 인공지능망에 대한 입력 값과 출력 값만 알고 있는 상태에서 RGF (Random Gradient Free) 알고리즘을 이용하여 섭동을 최적화함으로써, 블랙박스 적대적 공격 중에서도 의사결정 기반 적대적 공격을 수행할 수 있게 된다.

도면의 간단한 설명

- [0033] 도 1는 예시적인 실시예들에서 사용되기에 적합한 컴퓨팅 장치를 포함하는 컴퓨팅 환경을 예시하여 설명하기 위한 블록도
- 도 2는 본 발명의 일 실시예에 따른 인공 신경망 모델의 적대적 공격 방법을 설명하기 위한 흐름도
- 도 3은 본 발명의 일 실시예에 따른 인공 신경망 모델의 적대적 공격 방법에서 섭동에 대한 최적화를 수행하는 과정을 나타낸 도면
- 도 4는 본 발명의 일 실시예에 따른 S220을 상세히 설명하기 위한 흐름도

발명을 실시하기 위한 구체적인 내용

- [0034] 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.
- [0035] 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.
- [0036] 이하의 설명에 있어서, 신호 또는 정보의 "전송", "통신", "송신", "수신" 기타 이와 유사한 의미의 용어는 일 구성요소에서 다른 구성요소로 신호 또는 정보가 직접 전달되는 것뿐만이 아니라 다른 구성요소를 거쳐 전달되는 것도 포함한다. 특히 신호 또는 정보를 일 구성요소로 "전송" 또는 "송신"한다는 것은 그 신호 또는 정보의 최종 목적지를 지시하는 것이고 직접적인 목적지를 의미하는 것이 아니다. 이는 신호 또는 정보의 "수신"에 있어서도 동일하다. 또한 본 명세서에 있어서, 2 이상의 데이터 또는 정보가 "관련"된다는 것은 하나의 데이터(또는 정보)를 획득하면, 그에 기초하여 다른 데이터(또는 정보)의 적어도 일부를 획득할 수 있음을 의미한다.
- [0037] 한편, 본 발명의 실시예는 본 명세서에서 기술한 방법들을 컴퓨터상에서 수행하기 위한 프로그램, 및 상기 프로그램을 포함하는 컴퓨터 판독 가능 기록매체를 포함할 수 있다. 상기 컴퓨터 판독 가능 기록매체는 프로그램 명령, 로컬 데이터 파일, 로컬 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체는 본 발명을 위하여 특별히 설계되고 구성된 것들이거나, 또는 컴퓨터 소프트웨어 분야에서 통상적으로 사용 가능한 것일 수 있다. 컴퓨터 판독 가능 기록매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체, CD-ROM, DVD와 같은 광 기록 매체, 및 롬, 램, 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 상기 프로그램의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함할 수 있다.

- [0038] 도 1은 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.
- [0039] 도시된 컴퓨팅 환경(10)은 컴퓨팅 장치(12)를 포함한다. 일 실시예에서, 컴퓨팅 장치(12)는 본 발명의 실시예에 따른 인공 신경망 모델의 적대적 공격을 수행하기 위한 장치일 수 있다.
- [0040] 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.
- [0041] 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.
- [0042] 통신 버스(18)는 프로세서(14), 컴퓨터 판독 가능 저장 매체(16)를 포함하여 컴퓨팅 장치(12)의 다른 다양한 컴포넌트들을 상호 연결한다.
- [0043] 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.
- [0044] 도 2는 본 발명의 일 실시예에 따른 인공 신경망 모델의 적대적 공격 방법을 설명하기 위한 흐름도이다. 전술한 바와 같이, 본 발명의 일 실시예에 따른 인공 신경망 모델의 적대적 공격 방법은 하나 이상의 프로세서들, 및 상기 하나 이상의 프로세서들에 의해 실행되는 하나 이상의 프로그램들을 저장하는 메모리를 구비한 컴퓨팅 장치(12)에서 수행될 수 있다. 이를 위하여, 상기 인공 신경망 모델의 적대적 공격 방법은 하나 이상의 컴퓨터 실행 가능 명령어를 포함하는 프로그램 내지 소프트웨어의 형태로 구현되어 상기 메모리상에 저장될 수 있다.
- [0045] 또한, 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.
- [0046] 우선, 컴퓨팅 장치(12)는 인공신경망에 대하여 적대적 공격을 수행하기 위한 섭동을 생성할 수 있다(S210). 구체적으로, 컴퓨팅 장치(12)는 원본 데이터 세트에 포함된 이미지와 동일한 크기의 섭동을 생성하고, 생성된 섭동을 초기화할 수 있다. 여기서, 섭동은 인공신경망에 대하여 적대적 공격을 수행하기 위해 원본 데이터 세트에 추가하는 노이즈일 수 있다. 본 발명에서는 이러한 섭동을 최적화하여 인공신경망에 대하여 높은 공격 성공률과 상대적으로 적은 자원으로도 공격을 수행할 수 있는 적대적 공격 방법을 제공할 수 있다. 예를 들어, 섭동은 원본 데이터 세트에 포함된 이미지와 동일한 크기를 가질 수 있으며, 초기 섭동은 영행렬(섭동의 각 픽셀이 0)로 초기화할 수 있다. 이는 초기 섭동에 의한 영향력을 최소화하기 위한 것이다.
- [0047] 다음으로, 컴퓨팅 장치(12)는 생성된 섭동에 대한 최적화를 수행할 수 있다(S220).
- [0048] 도 3은 본 발명의 일 실시예에 따른 인공 신경망 모델의 적대적 공격 방법에서 섭동에 대한 최적화를 수행하는

과정을 나타낸 도면이다.

[0049] 도 3과 같이, 컴퓨팅 장치(12)는 원본 데이터 세트에 섭동을 추가하여 제1 적대적 데이터 세트를 생성하고, 원본 데이터 세트에 섭동 및 노이즈를 추가하여 제2 적대적 데이터 세트를 생성할 수 있다. 또한, 컴퓨팅 장치(12)는 제1 적대적 데이터 세트 및 제2 적대적 데이터 세트를 인공 신경망에 입력하고, 인공 신경망의 출력 값을 기반으로 RGF(Random Gradient Free) 알고리즘을 이용하여 섭동에 대한 기울기(Gradient)를 산출하며, 산출된 기울기 및 노이즈를 섭동에 반영하여 업데이트할 수 있다. 이 때, 컴퓨팅 장치(12)는 기 설정된 횟수 또는 기 설정된 기준이 될 때까지, 최적화를 수행하는 과정을 반복할 수 있다.

[0050] 도 4는 본 발명의 일 실시예에 따른 S220을 상세히 설명하기 위한 흐름도이다.

[0051] 도 4를 참조하면, 컴퓨팅 장치(12)는 원본 데이터 세트에 섭동을 추가하여 제1 적대적 데이터 세트를 생성할 수 있다. 여기서, 원본 데이터 세트가 배치 사이즈(Batch size, 예를 들어, 다수의 이미지)로 구성될 수 있다.

[0052] 다음으로, 컴퓨팅 장치(12)는 원본 데이터 세트에 섭동 및 노이즈를 추가하여 제2 적대적 데이터 세트를 생성할 수 있다. 여기서, 노이즈는 RGF(Random Gradient Free) 알고리즘을 이용하기 위하여 랜덤하게 부여하는 노이즈로써, 예를 들어, 가우시안 노이즈일 수 있다.

[0053] 다음으로, 컴퓨팅 장치(12)는 제1 적대적 데이터 세트 및 제2 적대적 데이터 세트를 인공 신경망에 입력할 수 있다.

[0054] 다음으로, 컴퓨팅 장치(12)는 인공 신경망의 출력 값을 기반으로 손실함수(loss function)를 이용하여 제1 적대적 데이터 세트 및 제2 적대적 데이터 세트에 대한 성능 지표를 각각 산출할 수 있다. 구체적으로, 컴퓨팅 장치(12)는 제1 적대적 데이터 세트에 대한 인공 신경망의 출력 값을 기반으로 손실함수를 이용하여 제1 적대적 데이터 세트에 대한 제1 성능 지표를 산출할 수 있다. 또한, 컴퓨팅 장치(12)는 제2 적대적 데이터 세트에 대한 인공 신경망의 출력 값을 기반으로 손실함수를 이용하여 제2 적대적 데이터 세트에 대한 제2 성능 지표를 산출할 수 있다. 이 때, 성능 지표는 데이터 세트에 대한 정확도 손실 값(점수)을 나타내는 것으로, 전체 데이터 세트 수에 대한 오답 데이터 수를 나타낼 수 있다. 정답 데이터는 원본 데이터 세트에 포함된 각 이미지에 부여된 라벨(label)을 통하여 알 수 있다. 예를 들어, 100장의 원본 데이터 세트에서 60장이 오답이면 0.6과 같이 성능 지표를 나타낼 수 있다.

[0055] 한편, 본 발명에서는 블랙박스 적대적 공격 중에서도 의사결정 기반 적대적 공격을 수행하기 위해 인공신경망에 대한 입력 값과 출력 값만 알고 있는 상태로, RGF 알고리즘을 이용하여 섭동을 최적화하고 있다. 이러한 RGF 알고리즘을 적용하려고 하는 한 손실 함수는 로컬 최소값에 빠지지 않도록 블록해야 하고, 안정적으로 최적화되기 위해서는 립시츠 연속 함수 (Lipschitz continuity)이어야 한다. 이에 본 발명에서는 손실함수으로써, 해밍 거리(hamming-distance)를 이용한다. 해밍 거리는 이진 요소로 구성된 두 배열 사이의 거리 측정법으로, 하기 수학식 1과 같이 정의할 수 있다.

[0056] [수학식 1]

$$L_H(\mathbf{y}_0, \mathbf{y}_1) = \frac{1}{N_{cls}} \sum_{i=1}^{N_{cls}} H(y_{0,i}, y_{1,i})$$

[0057]

[0058] (여기서, $L_H(\mathbf{y}_0, \mathbf{y}_1)$ 는 제1 성능 지표(손실 함수), $\mathbf{y}_0$ 는 제1 적대적 데이터 세트에 대한 인공 신경망의 출력 값, $\mathbf{y}_1$ 는 원본 데이터 세트에 대한 인공 신경망의 출력 값, N_{cls} 는 출력될 수 있는 클래스(출력 값)의 개수, i 는 인공 신경망에서 출력되는 클래스의 각 번호, $H(y_{0,i}, y_{1,i})$ 는 각 i 에 대한 해밍 거리(Hamming distance))

[0059] 이 때, 해밍 거리에 의해 출력된 정확도 손실은 불연속적인 거리 측정법으로, 이를 손실 함수로 이용하기 위하여 연속 공간에서 표현하면 하기 수학식 2와 같이 정의할 수 있다.

[0060] [수학식 2]

$$L_H(\mathbf{y}_0, \mathbf{y}_1) = \frac{1}{N_{cls}} \sum_{i=1}^{N_{cls}} \begin{cases} 0, & \text{if } y_{0,i} + y_{1,i} < 1 \\ y_{0,i} + y_{1,i} - 1, & \text{otherwise} \end{cases}$$

[0061]

$$= \frac{1}{N_{\text{cls}}} \sum_{i=1}^{N_{\text{cls}}} \max(0, y_{0,i} + y_{1,i} - 1)$$

$$L_H(\mathbf{y}_u, \mathbf{y}_1) = \frac{1}{N_{\text{cls}}} \sum_{i=1}^{N_{\text{cls}}} \max(0, y_{u,i} + y_{1,i} - 1)$$

(여기서, $L_H(\mathbf{y}_0, \mathbf{y}_1)$ 는 제1 성능 지표, $L_H(\mathbf{y}_u, \mathbf{y}_1)$ 는 제2 성능 지표, $\mathbf{y}_0$ 는 제1 적대적 데이터 세트에 대한 인공 신경망의 출력 값, $\mathbf{y}_u$ 는 제2 적대적 데이터 세트에 대한 인공 신경망의 출력 값, $\mathbf{y}_1$ 는 원본 데이터 세트에 대한 인공 신경망의 출력 값, N_{cls} 는 출력될 수 있는 클래스(출력 값)의 개수, i 는 인공 신경망에서 출력되는 클래스의 각 번호)

한편, 본 발명에서는 성능 지표를 정확도 손실 값, 즉 점수로 표현하여 산출하였으나, 이에 한정되는 것은 아니며, 성능 지표를 히스토그램으로 표현함으로써, EMD(Earth Moving Distance)와 Cross-Entropy를 이용하여 산출할 수 있다.

다음으로, 컴퓨팅 장치(12)는 제1 성능 지표 및 제2 성능 지표를 기반으로 RGF(Random Gradient Free) 알고리즘을 이용하여 섭동에 대한 기울기(Gradient)를 산출할 수 있다. 즉, 컴퓨팅 장치(12)는 하기 수학적 식 3과 같이, 섭동에 대한 기울기를 산출할 수 있다.

[수학적 식 3]

$$g = \frac{L_H(\hat{\mathbf{y}}_u, \mathbf{y}) - L_H(\hat{\mathbf{y}}_0, \mathbf{y})}{\gamma}$$

(여기서, g 는 섭동에 대한 기울기, $L_H(\hat{\mathbf{y}}_0, \mathbf{y})$ 는 제1 성능 지표, $L_H(\hat{\mathbf{y}}_u, \mathbf{y})$ 는 제2 성능 지표, γ 는 기울기 계산을 위한 상수, $\hat{\mathbf{y}}_u$ 는 제2 적대적 데이터 세트에 대한 인공 신경망의 출력 값, $\hat{\mathbf{y}}_0$ 는 제1 적대적 데이터 세트에 대한 인공 신경망의 출력 값, $\mathbf{y}$ 는 원본 데이터 세트에 대한 인공 신경망의 출력 값)

다음으로, 컴퓨팅 장치(12)는 섭동에 대한 기울기 및 노이즈를 섭동에 반영하여 섭동을 업데이트할 수 있다. 즉, 컴퓨팅 장치(12)는 하기 수학적 식 4와 같이, 섭동을 업데이트할 수 있다.

[수학적 식 4]

$$p_{t+1} = p_t + \alpha \cdot u_t \cdot \frac{L_H(\hat{\mathbf{y}}_{u_t}, \mathbf{y}) - L_H(\hat{\mathbf{y}}_0, \mathbf{y})}{\gamma}$$

(여기서, p_{t+1} 는 업데이트된 섭동, p_t 는 t 번째 섭동, α 는 업데이트 단계에 따른 크기 값, u_t 는 t 번째 노이즈)

이 때, 컴퓨팅 장치(12)는 업데이트 시, 섭동에 대한 기울기에 따라 t 번째 노이즈의 방향을 설정할 수 있다. 예를 들어, 제2 성능 지표가 제1 성능 지표보다 큰 경우, t 번째 노이즈를 양의 방향으로 설정할 수 있으며, 제2 성능 지표가 제1 성능 지표와 같거나 작은 경우, t 번째 노이즈를 음의 방향으로 설정할 수 있다.

한편, 컴퓨팅 장치(12)는 섭동의 크기가 너무 커지지 않도록 크기를 제한할 수 있다. 예를 들어, 섭동 중 크기가 기 설정된 기준을 초과하는 값이 있는 경우, 해당 값을 기 설정된 기준 값으로 변경함으로써, 크기를 제한할 수 있다.

또한, 컴퓨팅 장치(12)는 기 설정된 횟수 또는 기 설정된 기준이 될 때까지, S221 내지 S226을 반복 수행할 수 있다.

다음으로, 컴퓨팅 장치(12)는 최적화된 섭동을 이용하여 인공신경망에 대해 적대적 공격을 수행할 수 있다 (S230). 구체적으로, 컴퓨팅 장치(12)는 최적화된 섭동을 원본 데이터 세트에 추가하여 인공신경망에 대해 적대적 공격을 수행할 수 있다. 또한, 컴퓨팅 장치(12)는 적대적 공격의 수행 결과를 확인하여 리포트를 작성할 수 있다.

따라서, 본 발명의 일 실시예에 따른 인공 신경망 모델의 적대적 공격 방법은 원본 데이터 세트에 추가되는 섭

동을 최적화함으로써, 높은 공격 성공률과 상대적으로 적은 자원으로도 적대적 공격을 수행할 수 있게 된다.

[0079]

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

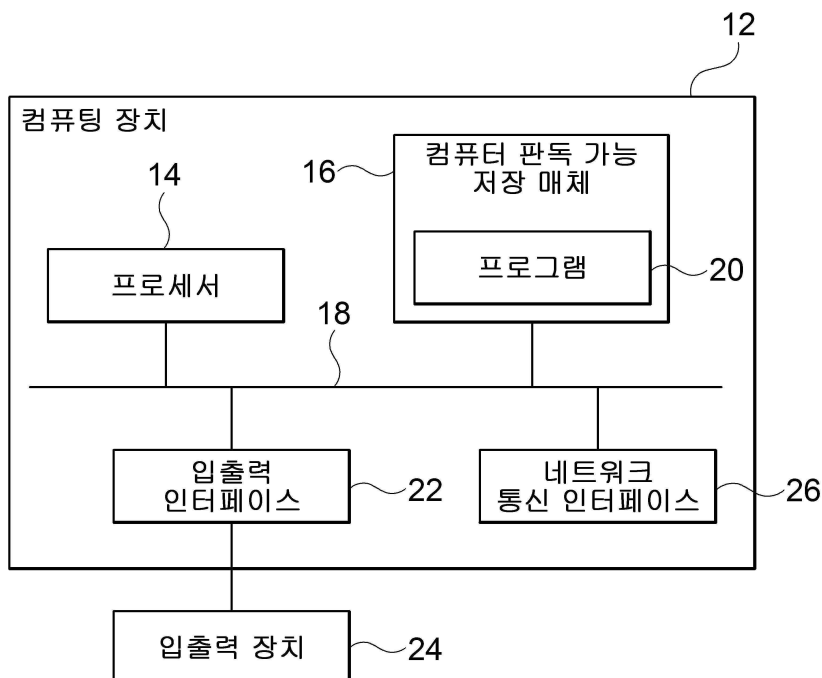
[0081]

- 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

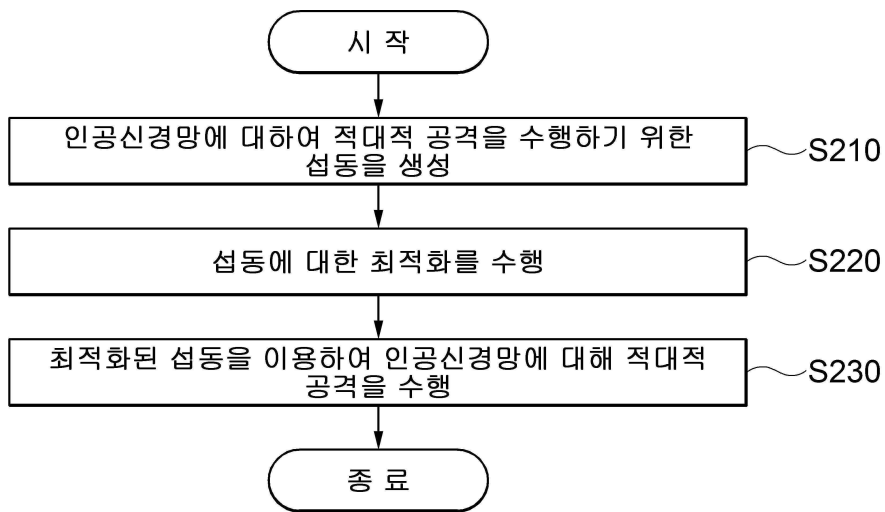
도면

도면1

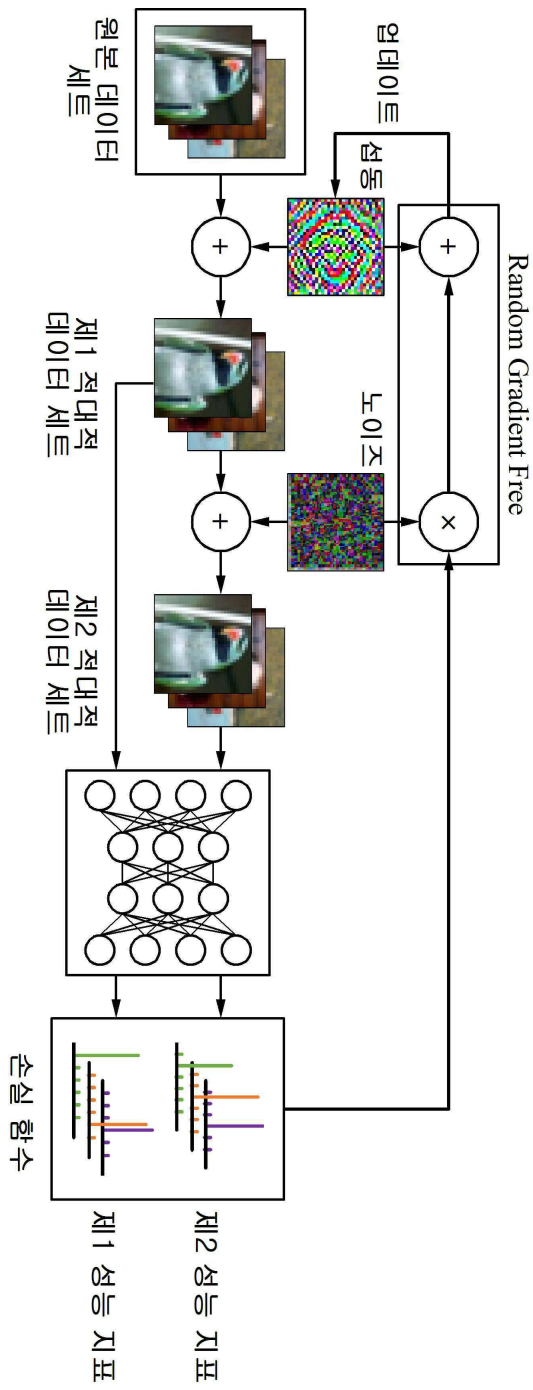
10



도면2



도면3



도면4

