(12) 특허협력조약에 의하여 공개된 국제출원

(19) 세계지식재산권기구
국제사무국

(43) 국제공개일
2023 년 1 월 5 일 (05.01.2023)

WIPO | PCT

(10) 국제공개번호
WO 2023/277206 A1

(51) 국제특허분류:
G06F 21/55 (2013.01) *G06N 3/08* (2006.01)

(21) 국제출원번호: PCT/KR2021/008127

(22) 국제출원일: 2021 년 6 월 28 일 (28.06.2021)

(25) 출원언어: 한국어

(26) 공개언어: 한국어

(30) 우선권정보:
10-2021-0083908 2021 년 6 월 28 일 (28.06.2021) KR

(71) 출원인: 동서대학교 산학협력단 (**DONGSEO TECHNOLOGY HEADQUTERS**) [KR/KR]; 47011 부산시 사상구 주례로 47, Busan (KR).

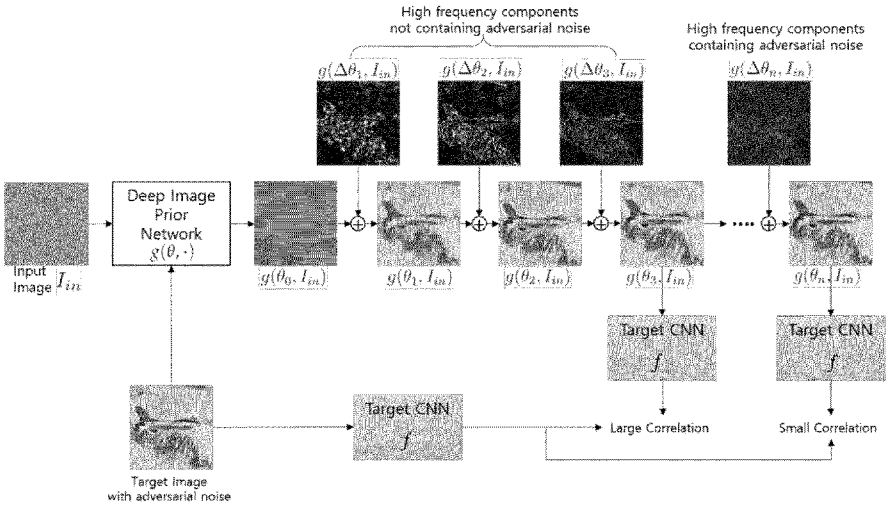
(72) 발명자: 이석호 (**LEE, Suk Ho**); 48439 부산시 남구 진남로25번길 13, 402, Busan (KR). 리차드수탄토 (**RICHARD, Evan Sutanto**); 47010 부산시 사상구 진사로36번가길 8-4, Busan (KR).

(74) 대리인: 김석계 (**KIM, Seok Kea**); 50918 경상남도 김해시 분성로 259, 3층 김석계특허법률사무소, Gyeongsangnam-do (KR).

(81) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의 국내 권리의 보호를 위하여): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(54) Title: SYSTEM FOR PREVENTING AI DYSFUNCTION CAUSED BY DECEPTION ATTACK

(54) 발명의 명칭: 기만공격에 의한 AI 역기능 방지 시스템



(57) Abstract: The present invention relates to a system for preventing AI dysfunction caused by a deception attack and, more specifically, is to construct an AI system which is robust to an AI deception attack on the basis of a deep image prior (DIP) network which detects adversarial noise, and relates to: a technology which can detect a base technology for vulnerable points in an AI system which is vulnerable to an adversarial example, and detect in advance whether an AI deception attack is being made; and a design of an AI system which is robust to such an attack. Accordingly, the present invention relates to a system for preventing AI dysfunction caused by a deception attack, wherein the system presents a standard guideline or an evaluation guideline which allows AI system designers to design a stable system, and applies the standard guideline or the evaluation guideline to an AI system certification management platform service system.

(57) 요약서: 본 발명은 기만공격에 의한 AI 역기능 방지 시스템에 관한 것으로, 보다 상세하게는 적대적 노이즈를 탐지하는 Deep Image Prior (DIP) 네트워크를 기반으로 하여 AI 기만공격(deception attack)에 강건한(robust) AI 시스템 구축을 하는 것으로, 적대적 사례(adversarial example)에 취약한 AI 시스템의 허점에 대한 기반 기술과 사전에 AI 기만 공격이 이루어지고 있는지를 사전에 감지할 수 있는 기술, 그리고 이러한 공격에 대하여 강건한 AI 시스템의 설계에 대한 것이다. 이를 통하여, AI 시스템 설계자들이 안정적인 시스템을 설계할 수 있는 표준지침 또는 평가 가이드라인을 제시하고 이를 AI 시스템 인증관리 플랫폼 서비스 시스템에 적용하는 기만공격에 의한 AI 역기능 방지 시스템에 관한 것이다.

WO 2023/277206 A1

[다음 쪽 계속]

(84) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의 역내 권리의 보호를 위하여): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 유라시아 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 유럽 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

공개:

— 국제조사보고서와 함께 (조약 제21조(3))

명세서

발명의 명칭: 기만공격에 의한 AI 역기능 방지 시스템

기술분야

[1] 본발명은 기만공격에 의한 AI 역기능 방지 시스템에 관한 것으로, 보다 상세하게는 AI기만공격(deception attack)에 강건한(robust) AI 시스템 구축을 하는 것으로, 적대적 사례(adversarial example)에 취약한 AI 시스템의 허점에 대한 기반 기술과 사전에 AI기만 공격이 이루어지고 있는지를 사전에 감지할 수 있는 기술, 그리고 이러한 공격에 대하여 강건한 AI시스템의 설계에 대한 것이다. 이를 통하여, AI 시스템 설계자들이 안정적인 시스템을 설계할 수 있는 표준지침 또는 평가 가이드라인을 제시하고 이를 AI 시스템 인증 관리 플랫폼 서비스 시스템에 적용하는 기만공격에 의한 AI 역기능 방지 시스템에 관한 것이다.

배경기술

[2] 최근 AI(Artificial Intelligence:인공지능)기술은 제조, 금융, 의료 등 다양한 사업 분야로 점차로 확대가 되고, 4차 산업혁명을 주도할 것으로 예상되고 있다. 그러나 AI에 대한 경제적, 사회적 의존도가 커질수록 AI시스템이 오동작하였을 때 감당해야 할 비용 또한 그만큼 천문학적으로 늘어나게 될 것이다. 즉, AI 시스템의 잘못된 설계나 사전에 설계단계에서 감지하지 못한 AI허점들은 인류발전에 역효과를 가져올 수 있다. 그러므로 최근 AI기반의 애플리케이션 개발만큼이나 AI기반의 시스템의 안정성 등에 대한 연구가 큰 관심을 끌고 있다.

[3] 최근 여러 연구에 따르면 인공 지능 (AI) 시스템은 정상적인 채널을 통해 들어오는 의도적으로 조작 된 데이터로 인해 오작동 할 수 있다. 이러한 종류의 조작 된 데이터를 적대적 예라고 한다. 적대적 사례는 공격자가 적대적 공격이라고하는 AI 시스템을 공격하는 수단으로 사용하는 경우 AI 주도 사회에 큰 위협이 될 수 있다. 이에 구글과 같은 주요 IT 기업들은 효과적인 방어 방법을 개발하여 적대적 공격에 강한 AI 시스템을 구축하는 방법을 연구하고 있다. 하지만 효과적인 방어 체제 구축이 어려운 이유 중 하나는 상대방이 어떤 적대적 공격 방식을 사용하고 있는지 미리 알기 어렵다는 점 때문이다.

[4] 4 차 산업 혁명을 선도하는 기술 중 인공 지능 (AI)에 대한 관심이 높아지고 있으며 자율 주행 차, 드론, 로봇 등 다양한 분야에 적용하여 그 기반을 확대하고 있다. 또한 과거 실험실 수준에서만 진행되던 AI 기반 자율 주행 차에 대한 연구는 이미 상용화 단계에 이르렀고, 구글, 애플, 삼성 등 IT 기업이 적극적으로 상용화 연구를 진행하고있다. 또한 제조, 금융, 헬스 케어 등 다양한 사업 분야에 기여하는 AI 기술은 사회 전반에 광범위한 영향과 경제적 파급 효과를 창출하고 있다. 그러나 인공 지능에 대한 사회적 의존도가 높아짐에 따라 인공 지능 오작동으로 인한 사회적 비용도 증가 할 것으로 예상된다.잘 설계된 소량의 노이즈가 호출되면 AI가 더 이상 이미지를 올바르게 인식하지 못하는 것으로

나타났다. 적대적인 노이즈가 이미지에 추가된다. 적대적 노이즈와 이미지의 조합을 적대적 예라고 한다. 이러한 적대적인 사례를 사용하여 의도적으로 AI 시스템을 오작동시키는 것을 AI기만 공격이라고 한다. 이러한 기만 공격의 심각성은 비정상적인 경로를 통해 시스템에 침입하지 않고 기만 공격을 할 수 있다는 점에 있다. 즉, 생성 엔진을 사용하여 공격자를 속이는 등 해킹으로부터 100 % 보호되는 AI 시스템조차도 일반 채널을 통해 들어오는 적대적 사례에 의해 최대한 공격받을 수 있다.

[5] 방어 방법은 AI 시스템이 denoised 적대적 사례를 시스템에 입력 할 수 있는지 여부에 따라 두 가지 범주로 나뉜다. denoised 적대적 예제의 입력을 허용하는 방어 방법은 denoised 적대적 예제가 더 이상 시스템에 해를 끼칠 수 없다고 가정한다. 이러한 방어 방법에는 AI 모델을 적대적 공격에 덜 취약하게 만드는 일종의 적대적 훈련 기법 인 방어 증류 기술과 난독화 된 기울기를 사용하여 공격자가 적대적 사례를 생성하기 위해 실행 가능한 기울기를 계산하기 어렵게 만드는 방법이 포함된다.

발명의 상세한 설명

기술적 과제

[6] 따라서 본발명은 상기와 같은 문제점을 해결하고자 안출된 것으로, 공격자가 사용하는 적대적 잡음의 종류에 대한 지식없이 적대적 잡음을 탐지하는 방법을 제안한다. 이를 위해 오직 다음으로 만 훈련 된 블러링 네트워크를 제안한다. 딥 이미지 사전 (DIP) 네트워크의 초기 조건으로도 사용한다. 이것은 신경망 훈련을 위해 많은 적대적 잡음 이미지를 사용해야하는 다른 신경망 기반 탐지 방법과는 대조적이다.

[7] 본 발명에서는 AI기만공격(deception attack)에 강건한(robust) AI 시스템 구축을 연구개발 목표로 한다. 이를 위해서 적대적 사례(adversarial example)에 취약한 AI 시스템의 허점에 대한 기반 기술과 사전에 AI기만 공격이 이루어지고 있는지를 사전에 감지할 수 있는 기술, 그리고 이러한 공격에 대하여 강건한 AI시스템을 제안한다. 이를 통하여, AI 시스템 설계자들이 안정적인 시스템을 설계할 수 있는 표준지침 또는 평가 가이드라인을 제시한다. 그리고 이를 AI 시스템 인증 관리 플랫폼 서비스 시스템에 적용하여 실제 제품에 적용하고자 하는 것이다.

[8] 본 발명에서는 공격자가 어떤 종류의 적대적 잡음을 사용하는지 모른 채 적대적 잡음을 탐지하는 방법을 제안한다. 이를 위해 정상적인 무소음 이미지에 대해서만 학습 된 Deep Image Prior (DIP) 네트워크의 초기 조건으로 블러링 네트워크를 사용하여 학습 과정에서 적대적 노이즈 이미지가 필요하지 않다.

과제 해결 수단

[9] 본발명은 기만공격에 의한 AI 역기능 방지 시스템에 관한 것으로, 적대적 노이즈를 탐지하는 Deep Image Prior 네트워크 (DIP)를 기반으로 하는 것을

특징으로 한다.

[10] 또한, 상기 DIP가 실시간으로 이미지를 생성하기 위해 블러링 네트워크의 매개 변수를 초기 조건인 초기 매개 변수로 사용하는 것을 특징으로 한다.

[11] 또한, 상기 블러링 네트워크를 훈련시키는 기반 손실 함수를 사용하여 흐릿한 이미지가 올바른 분류 결과 방향으로 이미지를 흐리게하도록 훈련시키는 것을 특징으로 한다.

발명의 효과

[12] 본 발명의 결과를 통해 AI 시스템의 미검증된 위협요소를 제거할 수 있고, 보다 안정적인 AI 시스템이 제공되며, AI 시스템의 미검증된 위협요소를 제거할 수 있고, 보다 안정적인 AI 시스템 설계의 기초 사례, 기초 자료, 기반 기술 등을 제시함으로서 AI 시스템 설계자들은 안정적인 시스템을 설계할 수 있는 표준지침 또는 평가 가이드라인을 활용하여 개선된 AI시스템을 설계할 수 있고, AI 시스템 이용자들은 취약점이 제거된 안정적인 업그레이드된 AI시스템을 활용할 수 있는 효과가 있다.

[13] 또한 본 발명에서는 Deep Image Prior (DIP) 네트워크의 초기 조건으로 높은 수준의 표현 기반 블러링 네트워크를 사용하는 실시간 적대적 탐지 방법이므로, 고수준 표현 기반 블러링 네트워크는 정상적인 노이즈없는 이미지로만 훈련되며, 이는 신경망을 훈련하기 위해 많은 적대적 잡음 이미지를 사용해야하는 다른 신경망 기반 감지 방법과는 대조적이다. 네트워크 훈련을 위해 적대적 노이즈 이미지가 필요하지 않기 때문에 제안 된 방법은 공격 유형에 관계없이 AI기만 공격을 탐지 할 수 있다. 또한 본발명 모든 데이터 세트에서 다른 탐지 방법보다 우수한 것이며, 본발명 감지 방법은 실제 이미지에서도 작동하며 이미지 당 0.05 초 미만의 감지 속도를 보여준다. 이 성능은 본발명 방법이 실시간 AI 시스템에 적용 가능함을 보여준다. 본발명의 또 다른 큰 장점은 감지 방법이 결정 론적 신경망을 기반으로 하지 않고 들어오는 모든 이미지에 대해 매개 변수가 변경되는 이전의 깊은 이미지에 기반한다는 것이다. 이것은 공격자가 탐지를 피할 수 있는 적대적인 예를 생성하기 어렵게 만드는 현저한 효과가 있다.

도면의 간단한 설명

[14] 도 1. 인공 지능 (AI)기만 공격의 개념도.

[15] 도 2. 빠른 수렴을 위해 블러링 네트워크 매개 변수를 사용하는 DIP (Deep Image Prior) 초기화. (a) 느린 컨버전스의 기존 DIP (b) 블러링 네트워크 매개 변수로 초기화 된 DIP 도면.

[16] 도 3. 제안 된 방법의 개념도

[17] 도 4. 제안 된 AI기만 공격 탐지 방법 다이어그램.

[18] 도 5. 매개 변수 도메인에 대한 설명. (a) 블러링 네트워크를 통해 일반 이미지를 넣는 효과 (b) DIP를 통해 일반 이미지를 넣는 효과 (c) 블러링 네트워크를 통해

적대적인 이미지를 넣는 효과 (d) DIP를 통해 적대적인 이미지를 넣는 효과 도면
.

[19] 도 6. (a) 적대적 이미지의 상위 5 개 클래스에 대한 정확도 값 표시. (b) 입력으로 (a)를 사용하는 DIP의 출력에 대해 (a)에서와 동일한 5 개의 클래스. (c) 정상 이미지의 상위 5 개 클래스. (d) 입력으로 (c)를 사용하는 DIP의 출력에 대해 (c)에서와 동일한 5 개의 클래스 도면.

[20] 도 7. 제안된 방법의 순서도

발명의 실시를 위한 최선의 형태

[21] 본발명을 첨부도면에 의해 상세히 설명하면 다음과 같다.

[22] 본발명의 AI기만 공격의 개념에 대해 설명하면, 적대적 잡음은 신경망의 원래 입력에 추가 될 때 신경망이 잘못된 결정을 내릴 수 있도록 신중하게 설계된 작은 섭동이다. 원래 입력과 적대적 잡음의 조합을 적대적 예라고한다. 이러한 적대적인 예를 사용하여 의도적으로 AI 시스템을 오작동시키는 것을 AI기만 공격이라고 한다. 이러한 AI기만 공격의 심각성은 비정상적인 경로를 통해 시스템에 침입하지 않고 기만 공격을 할 수 있다는 점에 있다. 이는 비정상 경로를 통해 시스템에 침입하는 해킹 공격과는 대조적이다. 따라서 해킹 공격에 대해 안전한 AI 시스템조차도 정상적인 채널을 통해 들어오는 적대적 사례에 의해 공격받을 수 있다. 그림 1은 AI기만 공격의 개념을 보여준다. 이미지에 추가된 노이즈가 작아서 인간의 눈에는 원본 이미지와 적대적인 예가 비슷해 보이지만 신경망은 두 이미지에 대해 다른 결정을 내린다. 이런 종류의 적대적인 예는 신경망의 결정에 따라 시스템에 치명적인 피해를 줄 수 있다.

[23] 도 1. 인공 지능 (AI)기만 공격의 개념도는 원본 이미지에 작은 적대적 노이즈가 추가되면 신경망이 이미지를 이집트 고양이 대신 과카 몰리로 분류 할 수 있다. 이는 비정상적인 경로를 통해 시스템을 침입하는 해킹 공격과는 대조적이다.

[24] 적대적인 예를 생성하는 방법에는 여러 가지가 있다. 현재까지 가장 널리 사용되는 적대적 예제 생성 방법 중 하나는 FGSM (Fast Gradient Sign Method)이다. 이 방법은 그라디언트를 사용하여 신경망의 출력과 실제 레이블 사이의 거리를 늘려서 적대적인 예를 생성한다.

[25]

$$\hat{x} = x + \epsilon \cdot \text{sign}(\nabla_x J(x, y_{\text{true}})). \quad (1)$$

[26] 여기서 x는 입력 영상, x*?*는 생성 된 적대 영상, sign (a)는 a의 부호를 취하는 부호 연산자, ytrue는 입력 영상의 실제 레이블, ∇x는 x에 대한 기울기, 그리고 ε은 작은 양의 값이다. FGSM의 간단한 확장은 BIM (Basic Iterative Method) [5]으로 적대적 잡음 η을 작은 ε 값으로 반복적으로 적용한다.

[27]

$$x_0^* = x$$
$$x_i^* = \text{clip}_{x,\epsilon}\left(x_{i-1}^* + \epsilon\, \text{sign}\left(\nabla_{x_{i-1}^*} J(\Theta, x_{i-1}^*, y)\right)\right), \quad (2)$$

[28] 여기서 clipx, $\varepsilon(\cdot)$는 원본 이미지 x의 ε- 이웃 내에 놓이도록 적대적 예제의 값을 클리핑 한 것이다. 이 재귀 공식은 적대적 예제가 클래스 경계를 얼마나 멀리 넘어 가야하는지에 대한 제어를 제공하여 ImageNet 데이터 세트에 대한 FGSM 공격보다 더 효과적이라는 것을 보여준다.

[29] 가장 강력한 적대적 공격 중 하나는 최적화 기술을 사용하여 적대적 사례를 생성하는 Carlini-Wagner (CW) 공격 방법이다. Carlini-Wagner (CW) 공격 방법은 주어진 제약 조건을 최소화 함수로 이동하여 강력한 적대적 예를 생성함으로써 원래 최적화 문제를 재구성한다.

[30]

[31] 본 발명에서는 공격자가 사용하는 적대적 잡음의 종류에 대한 지식없이 적대적 잡음을 탐지하는 방법을 제안한다. 이를 위해 일반 이미지로만 훈련 된 블러링 네트워크를 제안하고 이를 DIP (Deep Image Prior) 네트워크의 초기 조건으로 사용한다. 이것은 신경망 훈련을 위해 많은 적대적 잡음 이미지를 사용해야하는 다른 신경망 기반 탐지 방법과는 대조적이다. DIP는 입력 이미지의 노이즈가 제거 된 버전을 재구성 한 다음 제안 된 감지 조치를 사용하여 원래 입력과 비교하여 입력이 적대적 노이즈인지 여부를 결정한다.본발명 방법은 실제 이미지에 대해 더 나은 노이즈 제거 특성을 가지며 더 높은 검출 정확도를 얻었다. 본발명은 적대적 잡음을 특성화하는 정교한 통계 도구 설계 나 공격자가 만든 적대적 공격 유형에 대한 지식을 요구하지 않는다. 또한 네트워크를 훈련시키기 위해 적대적인 예제가 필요하지 않다.

[32]

[33] 본발명은 DIP의 초기 조건으로 블러링 네트워크의 매개 변수를 사용하는 적대적 예제 탐지 방법이다.

[34] 본발명 방법의 주요 기술적 구성을 설명한다. 그림 3은 주요 기술의 개념도를 보여준다. 주요 기술은 타겟 CNN (Convolutional Neural Network)의 출력을 두 개의 다른 입력, 즉 적대적 노이즈를 포함하는 테스트 이미지 (Itest)와 DIP에 의해 재구성 된 이미지를 비교하는 것이다. DIP에 의해 재구성 될 대상으로 테스트 이미지를 사용하여 DIP는 DIP에 대한 노이즈 입력 (Iin)을 테스트 이미지로 천천히 재구성한다. 재구성 과정에서 적대적 잡음을 포함하지 않는 고주파 성분이 먼저 재구성되고 적대적 잡음은 나중에 재구성되는데, 이는 DIP의 잡음 저항 특성 때문이다. 따라서 적대적 잡음이 재구성되기 전에 재구성 프로세스가 중단되면 재구성 된 이미지는 대상 CNN에서 원래 테스트 이미지와 다른 효과를 나타냅니다. 따라서 타겟 CNN의 두 출력 간의 상관관계를 측정하여 입력에 적대적 잡음이 포함되어 있는지 여부를 확인할 수 있다.

[35]

[36] 그림 4는 본발명 방법의 전체적인 다이어그램을 보여준다. 테스트 시간에 입력 이미지 Iin은 깨끗한 이미지이거나 적대적인 이미지 일 수 있다. 감지는 서로 다른 입력에 대한 대상 모델 (타겟 컨볼 루션 신경망)의 출력을 비교하는 감지

측정 D를 기반으로 한다. 검출 측정 D는 입력 f (Iin)로 취하고 f (g (θ △θ, Iin)), 여기서 f (·)는 사전 훈련 된 분류기의 출력, 즉 탐지 방법으로 방어하려는 대상 모델을 나타냅니다. 또한, g (θ △, Iin) = gB (θ △, Iin)는 블러링 네트워크의 초기 매개 변수 θ에서 매개 변수가 △θ만큼 업데이트 된 후 DIP 네트워크 g (·)의 출력을 나타냅니다. 즉, g (θ, Iin) = gB (θ, Iin), Iin은 입력 영상이다.

[37]

[38] 블러링 네트워크 gB의 역할은 두 가지이다. 첫째, gB는 입력 이미지를 블러링하여 적대적인 노이즈를 제거한다. 둘째, gB의 매개 변수는 DIP (Deep Image Prior) 네트워크의 초기 조건 역할을한다. 다음 섹션에서는 먼저 블러링 네트워크로서 gB의 역할을 설명한다. 그런 다음 DIP의 초기 조건으로서 gB의 역할과 탐지 측정 D가 어떻게 정의되는지 설명한다.

[39]

[40] 블러링 네트워크로서 gB의 역할에 대해 설명하면 다음과 같다.

[41] 네트워크 gB는 입력 이미지를 흐리게하도록 훈련되었습니다. 블러링은 고주파 성분을 제거하므로 적대적인 노이즈를 어느 정도 제거한다. 그러나 고주파 성분을 제거하면 올바른 분류에 도움이 되는 성분도 제거된다. 따라서 이러한 원치 않는 부작용을 방지하기 위해 gB를 훈련 할 때 손실 함수에 높은 수준의 표현 안내 용어를 추가하여 gB가 높은 수준의 응답이 비 수준의 응답과 유사한 방향으로 이미지를 흐리게 할 것이다. 흐릿한 노이즈없는 이미지이다.

[42] 그림 5에서 제안 된 방법의 주요 아이디어를 설명하며, 이미지의 높은 수준의 특징 공간을 보여준다. 가로축은 블러 정도의 축이다. 즉, 특징 벡터가 더 오른쪽에 배치 될수록 더 블러링 된 이미지의 특징 벡터에 해당한다. 그림 5a, b에서 도메인 Otrue는 올바르게 분류 된 이미지의 도메인이다. 흐려짐이 심화됨에 따라 흐릿한 이미지에 해당하는 특징 벡터는 Otrue의 경계를 넘어서 잘못 분류된다. 이는 올바른 분류에 기여하는 고주파 성분이 블러링 프로세스에 의해 제거되기 때문이다. 그러나 우리는 이 네트워크에 의해 블러링 된 이미지가 여전히 올바른 클래스로 분류되도록 블러링 네트워크를 훈련시키고자 하는 것이다. 보다 구체적으로, 그림 5a와 관련하여, gB가 θ1 대신 매개 변수 θ2를 갖도록, 즉 f (Iin)에서 f (gB (θ2, Iin) 로의 매핑을 달성하도록 네트워크 gB를 훈련 시키길 원한다.) 여전히주는 Iin이 노이즈가없는 깨끗한 이미지 인 경우 분류 결과가 f (Iin)와 다른 f (gB (θ1, Iin))가 아닌 f (Iin)와 동일한 분류 결과이다. 이는 다음과 같은 손실 함수로 gB를 훈련하여 달성 할 수 있다.

[43]

[44] $$\mathcal{L}_{blur} = \|g_B(I_{in}) - G_\sigma * I_{in}\|_2^2 + \lambda \|f(g_B(I_{in})) - f(I_{in})\|_2^2, \quad (4)$$

[45]

[46] 여기서 Gs는 표준 편차가 s 인 가우스 커널을 나타내고 *는 컨볼 루션 연산자를 나타냅니다. 첫 번째 항 (‖gB (Iin) - Gs * Iin‖2)의 최소화는 입력 이미지의 흐릿한

버전 인 출력을 생성하는 네트워크를 달성하고 두 번째 항 (‖f ()의 최소화를 목표로한다. gB (Iin)) - f (Iin) ‖2)는 Iin과 유사한 분류 결과를 갖는 결과를 얻는 것을 목표로하며, λ는 두 항 간의 균형을 제어하는 양수 값이다. 여기에서 훈련에서는 Iin에 대해 노이즈가없는 깨끗한 이미지 만 사용하는 반면 테스트 시간에는 Iin이 노이즈가 있거나 노이즈가 없는 이미지가 될 수 있다는 점에 유의해야한다.

[47] 한편, Iin에 적대적 잡음이 포함 된 경우 블러링 네트워크 gB가 잡음을 효과적으로 제거하기를 원한다. 즉, 그림 5c와 관련하여 네트워크가 f (gB (θ1)에 대한 매핑이 아니라 분류 결과가 Iin과 다른 f (Iin)에서 f (gB (θ2, Iin)) 로의 매핑을 달성하기를 원한다. Iin)) Iin과 동일한 잘못된 분류 결과를 제공한다. Iin이 깨끗한 이미지 인 경우와 달리 gB가 입력 이미지와 다른 분류 결과를 사용하여 잡음이있는 이미지를 공간에 매핑하려고하는 이유는 gB가 깨끗한 이미지에서만 훈련되기 때문이다. 따라서 f (Iin)와 f (gB (θ, Iin))의 차이가 크면 입력 영상이 적대적 영상임을 나타내고 작은 차이는 정상적인 영상임을 나타내며 이미 적대적 잡음을 설계 할 수 있다. f (Iin)과 f (gB (θ, Iin)). 간단한 측정은 다음과 같다.

[48]

$$S(I_{in},\theta) = \frac{1}{|f(g_B(\theta, I_{in})) - f(I_{in})| + \alpha} \tag{5}$$

[49]

[50] f (Iin)과 f (gB (θ, Iin))이 유사하면 큰 값을 가지며, 다르면 작은 값을 가지며, 여기서 a는 0으로 나누지 않도록 작은 양의 값이다. 따라서 S (Iin, θ)가 작 으면 Iin에 적대적 잡음이 있다는 결론을 내릴 수 있다. 따라서 다음과 같이 사전 정의 된 임계 값 (Th)과 비교하여 적대적 잡음을 감지한다.

[51]

$$\begin{cases} \text{Detect Adversarial Attack} & \text{if } S(I_{in},\theta) \geq Th \\ \text{No Adversarial Attack} & \text{if } S(I_{in},\theta) < Th. \end{cases} \tag{6}$$

[52] 네트워크 gB는 이미 적대적 잡음을 제거하고 탐지 할 수 있는 능력을 가지고 있다.

[53] 그러나 노이즈 제거 및 감지 기능을 높이기 위해 다음 섹션에서와 같이 한 단계 더 나아간다.

[54] DIP의 초기 조건으로서 gB의 역할에 대해 설명하면 다음과 같다.

[55] MAML의 개념에 따르면, 블러링 네트워크의 매개 변수는 입력 이미지 재생성 작업에 대한 좋은 초기 매개 변수로 간주 될 수 있다. 이러한 매개 변수는 서로 다른 입력 이미지의 흐릿한 버전을 생성하는 방법을 학습하여 얻었기 때문이다. 블러링 네트워크의 매개 변수를 초기 매개 변수로 사용하여 블러링 네트워크를 입력 이미지에 맞게 미세 조정한다. 즉, 입력 이미지를 재현하는 DIP의 초기

조건으로 블러링 네트워크의 매개 변수를 사용한다. 이렇게하면 DIP가 매우 빠르게 수렴되어 입력 이미지의 재구성이 실시간으로 수행 될 수 있다. 그림 2b는이 개념을 보여준다. 블러링 네트워크의 매개 변수를 사용하여 DIP는 먼저 입력 이미지의 흐릿한 버전을 재현한다. 매개 변수가 업데이트됨에 따라 DIP는 실시간으로 블러링 된 이미지의 더 선명하고 선명한 버전을 생성한다. 따라서 블러링 네트워크가 테스트 이미지에 과적합된다. Itest = gB (θ, Iin). 이는 다음 손실 함수를 최소화하여 달성 할 수 있다.

[56]

$$\mathcal{L}_{fine} = \|g(\theta, I_{test}) - I_{in}\|_2^2. \quad (7)$$

[57]

[58] (4)의 Iin은 훈련 데이터 세트의 모든 이미지를 나타내며, (7)의 Itest는 테스트 이미지 만 나타낸다.

[59] 몇 번의 업데이트 후, L2 norm 의미에서 Iin에 가까운 g (θB △θ, Itest)를 얻지 만 headversarialnoise는 포함하지 않는다. g (θ △θ, Itest)가 분류 자 f에 입력되면 정확하거나 적어도 유사한 분류 결과를 제공한다. Iin이 정상적인 이미지라면 Iin에게. 그림 5b는 매개 변수의 업데이트에 의해 일반 영상의 분류 결과가 f (Itest)와 유사하게 되는 경우를 보여준다. 반대로 Iin이 적대적인 이미지 인 경우 DIP의 과적 합은 고주파 구성 요소도 복원한다. 그러나 흐릿한 이미지에서 시작하기 때문에 몇 번의 반복만으로는 적대적 노이즈를 완전히 복원하기에 충분하지 않으며, f (g (θ △θ, Itest))는 그림 5d와 같이 f (Iin)와 여전히 다르다. 따라서 f (Iin) 및 f (g (θ △θ, Itest))에 대해 유사성 측정을 사용하면 정상 이미지와 적대적 이미지를 구분할 수 있다. 다음 섹션에서는 유사성을 기반으로 제안된 탐지 조치에 대해 자세히 설명한다.

[60]

[61] AI기만 공격 탐지를 위한 제안 된 탐지 조치에 대해 설명하면 다음과 같다.

[62]

[63] S5는 벡터 f (Iin)에서 상위 5 개의 활성화 값을 제공하는 노드 집합을 나타내고 fn (Iin)은 S5에서 n 번째 노드의 활성화 된 값을 나타냅니다. 유사성 제안 된 방법에서 탐지 조치로 사용하는 조치는

[64]

$$S(I_{in}, \theta + \Delta\theta) = \sum_{n \in S_5} \frac{f_n(g(\theta + \Delta\theta, I_{test})) + f_n(I_{in})}{|f_n(g(\theta + \Delta\theta, I_{test})) - f_n(I_{in})| + \alpha}. \quad (8)$$

[65]

[66] 여기서는 (5)와 비교하여 fn (g (θ △θ, Itest)) fn (Iin)을 곱해 fn (g (θ △θ, Itest)) 또는 fn (Iin). 실제로 네트워크의 맨 마지막 계층을 사용하는 대신 두 번째에서 마지막 계층까지 사용한다. 혼란을 피하기 위해 여전히 f로 표시한다. 또한 f의 모든 노드 값을 사용하는 대신 노드 집합, 즉 S5 집합 만 사용한다. 그림 6은

적대적 이미지의 상위 5 개 가능한 클래스에 대한 정확도 값과 정규 및 재구성 된 이미지의 동일한 클래스에 대한 해당 정확도 값을 보여준다. 블러링 네트워크와 DIP가 잘못된 클래스의 정확도를 감소시키기 때문에 재구성 된 이미지에서 상위 5 개 등급의 적대적 이미지에 대한 정확도 값이 그림 6a, b에서 확인할 수 있다. 따라서 이 경우 S (Iin, θ △θ)가 작아진다. 반면에 입력이 정상 이미지이면 재구성 된 이미지는 입력 이미지의 상위 5 개 클래스에 대해 다소 유사한 정확도 값을 보여준다. 블러링 네트워크와 DIP가 일반 이미지와 동일한 분류 결과를 선호하는 방향으로 이미지를 재구성하기 때문이다. 따라서 미리 정의 된 임계 값을 사용하면 S (Iin, θ △θ) 값과 임계 값을 비교하여 AI기만 공격이 발생했는지 확인할 수 있다.

[67]

[68] 마지막으로, 탐지 측정에 대한 간단한 분석은 분류 결과가 다소 정확하지 않더라도 왜 우수한 탐지 성능을 보이는지 이해하는 데 도움이 된다. fn (g (θ △θ, Itest)) = fn (Iin) ϵn, 여기서 ϵn은 다음 값의 차이를 나타냅니다. fn (g (θ △θ, Itest)) 및 fn (Iin). 그런 다음 (8)을 다음과 같이 다시 작성할 수 있다.

[69]

$$S(I_{in},\theta) = \sum_{n\in S_5} \frac{|f_n(I_{in}) + \epsilon_n + f_n(I_{in})|}{|f_n(I_{in}) + \epsilon_n - f_n(I_{in})| + \alpha}. \quad (9)$$

$$= \sum_{n\in S_5} \frac{|2f_n(I_{in}) + \epsilon_n|}{|\epsilon_n| + \alpha}.$$

[70] ϵn 값은 다음과 같은 사실로 인해 음수이다. Iin이 적대적 이미지 인 경우 fn (g (θ △θ, Itest)) <fn (Iin) 적대적 이미지의 상위 5 개 클래스에 대한 정확도 값이 재구성 된 영상의 경우 Iin이 정상 영상이면 다시 fn (g (θ △θ, Itest)) <fn (Iin) 재구성 된 영상의 정확도 값은 해당 영상에 대한 일반 영상의 정확도 값만큼 클 수 없다. 정상 이미지의 상위 5 개 클래스. Iin이 적대적 이미지라면 | ϵn | Iin이 정상적인 이미지이면 | ϵn | 작다. 이는 Iin에 적대적 잡음이 포함 된 경우 g (θ, Iin)와 Iin이 크게 다르기 때문이다. g (θ △θ, ·)는 적대적 잡음의 잡음 제거기 역할을 하기 때문이다. 따라서 | 2fn (Iin) ϵn |에 대해 a를 무시한다. 1보다 작은 값을 가지려면 fn (Iin)은 | ϵn | a | ϵn |, Iin이 적대적 이미지라면 쉽게 그렇게 될 것이다. 따라서 정확도 값 fn (Iin)이 그다지 정확하지 않더라도 검출 성능이 좋다. 즉, DIP가 정확한 분류 결과를 제공하는 영상을 재구성하지 않더라도 검출 성능이 좋다. 이것이 제안 된 방법이 실험 섹션에서 볼 수 있듯이 좋은 탐지 성능을 보이는 이유 중 하나이다. 그림 7은 제안 된 방법의 순서도를 보여준다.

[71]

[72] 탐지 성능의 척도로 탐지의 정확도를 사용했다.

[73]

$$DetectionAccuracy = \frac{Number\,of\,correct\,detections}{Number\,of\,total\,detections} \times 100\%, \quad (10)$$

[74] 여기서 올바른 탐지는 탐지기가 입력 이미지를 적대적 예 또는 일반 이미지로 올바르게 분류했음을 의미한다. 테스트에 사용 된 적대적 예제 또는 일반 이미지의 수는 데이터 세트마다 다르다.

[75]

[76] 본발명 방법을 사용한 실험 설정에 대해 설명하면 다음과 같다.

[77] 블러링 네트워크와 DIP 네트워크 모두에 대해 스킵 연결이 있는 오토 인코더 아키텍처를 사용했다. 네트워크의 인코더 부분은 3 개의 컨볼 루션 레이어로 구성되며, 스트라이드 -2 다운 샘플링 컨볼 루션 연산에 따라 출력의 공간 크기가 감소한다. 인코더의 각 레이어에있는 필터 수는 각각 32, 64 및 128이다. 디코더는 또한 stride-2 디컨 볼 루션 연산에 따라 출력의 공간 크기가 증가하는 3 개의 컨벌루션 레이어로 구성된다. 디코더의 각 레이어에있는 필터의 수는 각각 128, 64 및 32이다. 두 네트워크 모두 네트워크의 모든 필터에 대해 -1과 1 사이의 균일 한 임의 초기화를 사용하는 AdamOptimizer로 훈련된다. 학습률을 0.0001로 고정하고 학습률 감소를 사용하지 않습니다. 과잉 훈련을 피하기 위해 10 회 연속 에포크 동안 검증 손실이 개선되지 않으면 블러링 네트워크를 조기에 중단했다. DIP를 사용하면 적대적인 노이즈가 아닌 이미지 만 복구하기 위해 조기 중지를 수행했다. 블러링 네트워크의 매개 변수를 DIP에 복사했을 때, 몇 번의 반복 만있는 DIP는 이미 위에서 설명한대로 이미지를 재구성했다. MNIST 및 CIFAR10 데이터 세트를 사용한 실험에는 10 회 반복을 사용하고 ImageNet 및 'Dog and Cat'데이터 세트를 사용한 실험에는 2 회 반복을 사용했다. 반복 횟수는 가장 높은 정확도를 제공하는 값으로 설정되며 많은 실험에서 수동으로 발견되었다. 이것은 (6)의 임계 값 (Th) 설정에도 적용된다. 수동으로 찾은 임계 값은 MNIST 및 CIFAR10 데이터 세트의 경우 98, ImageNet 및 'Dog and Cat'데이터 세트의 경우 74이다. (4)의 가우시안 커널 Gs에 대해 5 × 5 크기의 가우스 필터를 사용했다.

[78]

[79] MNIST 및 CIFAR10 데이터 세트의 결과에 대해 설명하면 다음과 같다.

[80]

[81] MNIST 및 CIFAR10 데이터 세트는 실제 이미지로 구성되어 있지 않지만 탐지 방법의 성능을 평가하기위한 참조 데이터 세트로 자주 사용된다. MNIST 및 CIFAR-10데이터 세트는 10 개의 서로 다른 클래스의 60,000 개 이미지로 구성된다. 여기서 CIFAR-10 데이터 세트의 이미지는 32 × 32 크기의 컬러 이미지이고 MNIST 데이터 세트의 이미지는 회색이다. 이미지 크기 28 × 28. FGSM 및 BIM 공격의 ε 값에 대해 각각 네 가지 설정을 사용하고 표 1에 표시된 CW 공격에 대해 두 가지 다른 신뢰 값 설정을 사용했다. MNIST 및 CIFAR10

데이터 세트 모두, 이미지의 절반 (5000 개 이미지)이 일반 이미지이고 나머지 절반 이미지 (5000 개 이미지)가 생성 된 적대적 노이즈와 함께 추가되었습니다. 대상 분류기는 5 개의 계층으로 구성된 단순 CNN (Convolutional Neural Network)이었습니다. 제안 된 방법은 실험과 동일한 DIP 구조를 사용한다. 'Dog and Cat'데이터 세트이지만 입력 및 출력 이미지의 크기는 CIFAR10 데이터 세트의 이미지 크기와 일치하며 입력 후 및 DIP 출력 전에 크기 조정 작업을 수행했다. 입력 이미지에 대해 적대적 노이즈가 생성되었으므로 크기 조정 작업은 제안 된 탐지기의 일부로 볼 수 있다.

[82]

[83] 표 1. 서로 다른 감지 방법 간의 감지 정확도 비교. 여기서 n' / w'는 '작동하지 않음'을,'-'는'알 수 없음'을 나타낸다.

[84] [표1]

Dataset	Attack Method	Strength	[18]	[26]	[27]	[31]	Proposed
MNIST	FGSM	$\epsilon = 0.1$	0.7768	0.7952	0.9514	0.8030	0.8566
		$\epsilon = 0.2$	0.8672	0.8977	0.9826	0.7767	0.6131
		$\epsilon = 0.3$	0.8925	0.9380	0.9887	0.8681	0.5613
		$\epsilon = 0.4$	0.9122	0.9424	0.9866	0.9246	0.5144
	BIM	$\epsilon = 0.1$	0.9419	0.8096	0.9716	0.8092	0.7894
		$\epsilon = 0.2$	0.9768	0.8330	0.9890	0.9027	0.5294
		$\epsilon = 0.3$	0.9801	0.7088	0.9896	0.9574	0.5039
		$\epsilon = 0.4$	0.9779	0.7002	0.9798	0.9797	0.4905
	CW	confidence = 0	n/w	0.9067	0.9419	0.6667	0.7058
		confidence = 10	n/w	0.9652	0.8893	0.6411	0.7926
CIFAR10	FGSM	$\epsilon = 1/255$	0.5521	0.5876	-	0.7108	0.7610
		$\epsilon = 3/255$	0.5892	0.5839	-	0.7097	0.8045
		$\epsilon = 6/255$	0.5994	0.6088	-	0.6667	0.8016
		$\epsilon = 9/255$	0.5931	0.6452	-	0.6671	0.8001
	BIM	$\epsilon = 1/255$	0.6588	0.5502	-	0.7807	0.7968
		$\epsilon = 3/255$	0.6172	0.6342	-	0.7755	0.8014
		$\epsilon = 6/255$	0.5962	0.8215	-	0.8117	0.8005
		$\epsilon = 9/255$	0.5802	0.8335	-	0.8654	0.7941
	CW	confidence = 0	n/w	0.8812	-	0.8757	0.7930
		confidence = 10	n/w	0.9489	-	0.8239	0.8705
Dog and Cat Dataset	FGSM	$\epsilon = 1$	n/w	n/w	0.9132	n/w	0.8516
		$\epsilon = 2$	n/w	n/w	0.6757	n/w	0.8847
		$\epsilon = 4$	n/w	n/w	0.8255	n/w	0.8908
		$\epsilon = 6$	n/w	n/w	0.5729	n/w	0.9096
	BIM	$\epsilon = 10$	n/w	n/w	0.8265	n/w	0.9155
		$\epsilon = 20$	n/w	n/w	0.7943	n/w	0.9010
	CW	confidence = 0	n/w	n/w	0.9198	n/w	0.9167
		confidence = 10	n/w	n/w	0.8365	n/w	0.9587
ImageNet	FGSM	$\epsilon = 1$	n/w	n/w	0.8643	n/w	0.8010
		$\epsilon = 2$	n/w	n/w	0.7763	n/w	0.8375
		$\epsilon = 4$	n/w	n/w	0.6116	n/w	0.8400
		$\epsilon = 6$	n/w	n/w	0.5026	n/w	0.8615
	BIM	$\epsilon = 10$	n/w	n/w	0.7789	n/w	0.8833
		$\epsilon = 20$	n/w	n/w	0.7431	n/w	0.8808
	CW	confidence = 0	n/w	n/w	0.9215	n/w	0.9243
		confidence = 10	n/w	n/w	0.8441	n/w	0.9650

[85] 표 1의 정확도 값에서 알 수 있듯이 본발명은 MNIST 데이터 세트의 이미지에 대한 FGSM 및 BIM 공격에 대한 다른 탐지 방법보다 나을 수 없다. 이는 MNIST 데이터 셋 소규모 그레이 스케일 이미지의 이미지에 복잡한 배경이 없고 주색(흑백)이 두 개뿐이어서 적을 식별 할 수 있는 이미지에서 통계적 특성을 쉽게 추출 할 수 있었기 때문이다. 본발명 방법을 사용하면 큰 흑백 적대적 소음도 쉽게 재구성 할 수 있다. DIP는 적대적 사건을 탐지하기 어려웠 기 때문에 제안된 방법의 정확도는 높지 않았다. 그러나 이러한 흑백 이미지는 실생활에서 흔하지 않기 때문에 MNIST 데이터 세트에 대한 실험은 ImageNet 또는 'Dog and Cat'데이터 세트에 대한 실험보다 덜 중요하다.

[86]

[87] 본 발명에서는 Deep Image Prior (DIP) 네트워크의 초기 조건으로 높은 수준의 표현 기반 블러링 네트워크를 사용하는 실시간 적대적 탐지 방법을 제안했다. 고수준 표현 기반 블러링 네트워크는 정상적인 노이즈없는 이미지로만 훈련되며, 이는 신경망을 훈련하기 위해 많은 적대적 잡음 이미지를 사용해야하는 다른 신경망 기반 감지 방법과는 대조적이다. 네트워크 훈련을 위해 적대적 노이즈 이미지가 필요하지 않기 때문에 제안 된 방법은 공격 유형에 관계없이 AI기만 공격을 탐지 할 수 있다. 또한 실험에서 제안 된 방법이 모든 데이터 세트에서 다른 탐지 방법보다 우수한 것으로 나타났다. 제안 된 감지 방법은 실제 이미지에서도 작동하며 이미지 당 0.05 초 미만의 감지 속도를 보여준다. 이 성능은 제안 된 방법이 실시간 AI 시스템에 적용 가능함을 보여준다. 본발명의 또 다른 큰 장점 방법은 감지 방법이 결정 론적 신경망을 기반으로하지 않고 들어오는 모든 이미지에 대해 매개 변수가 변경되는 이전의 깊은 이미지에 기반한다는 것이다. 이것은 제안 된 탐지 시스템을 알고있는 공격자가 탐지를 피할 수 있는 적대적인 예를 생성하기 어렵게 만든다.

청구범위

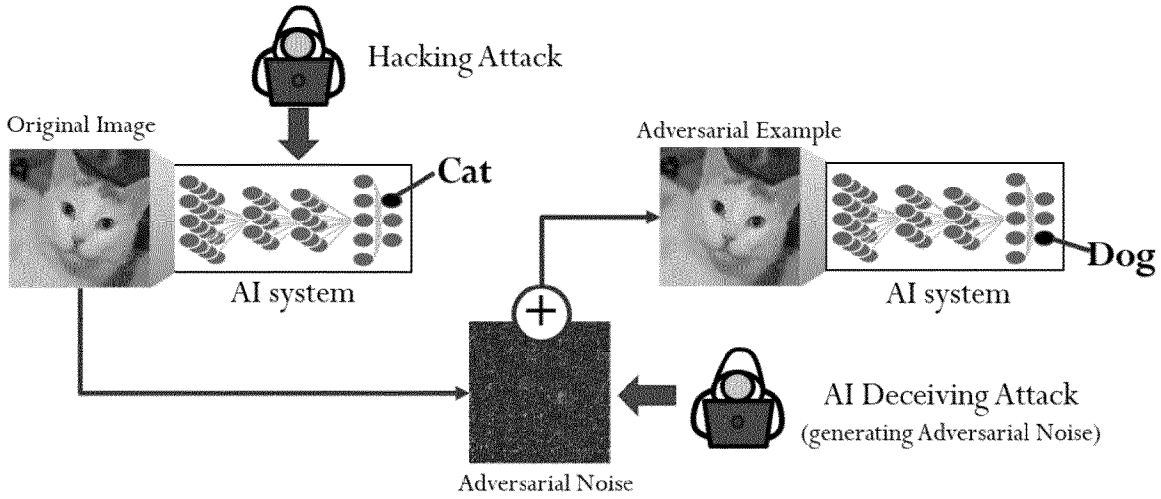
[청구항 1] 적대적 노이즈를 탐지하는 Deep Image Prior 네트워크 (DIP)를 기반으로 하는 것을 특징으로 하는 기만공격에 의한 AI 역기능 방지 시스템

[청구항 2] 제1항에 있어서, 상기 DIP가 실시간으로 이미지를 생성하기 위해 블러링 네트워크의 매개 변수를 초기 조건인 초기 매개 변수로 사용하는 것을 특징으로 하는 기만공격에 의한 AI 역기능 방지 시스템

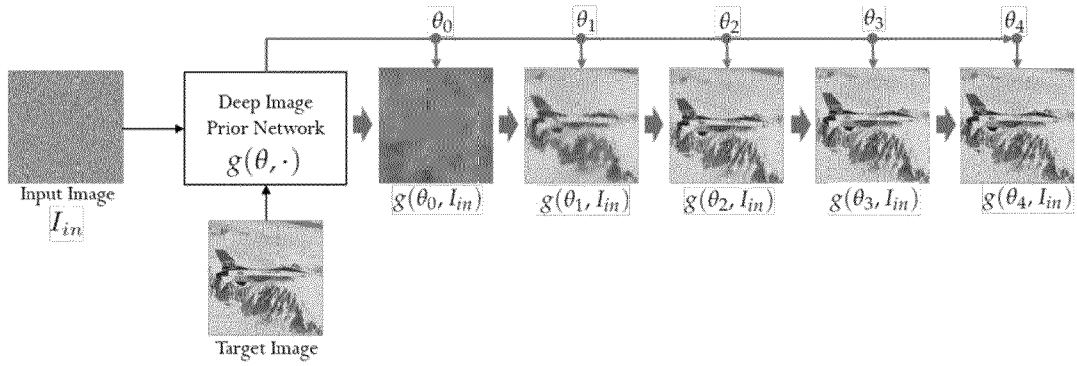
[청구항 3] 제2항에 있어서, 상기 블러링 네트워크를 훈련시키는 기반 손실 함수를 사용하여 흐릿한 이미지가 올바른 분류 결과 방향으로 이미지를 흐리게하도록 훈련시키는 것을 특징으로 하는 기만공격에 의한 AI 역기능 방지 시스템

WO 2023/277206 PCT/KR2021/008127

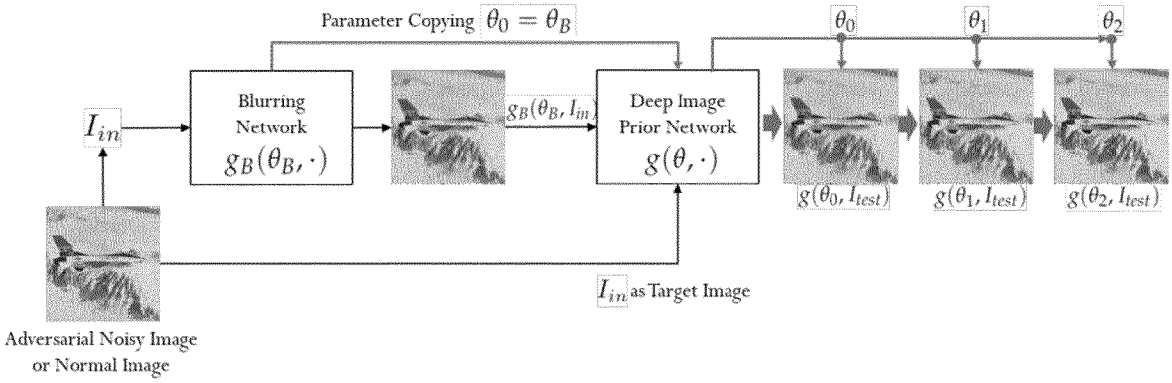
[도1]



[도2]

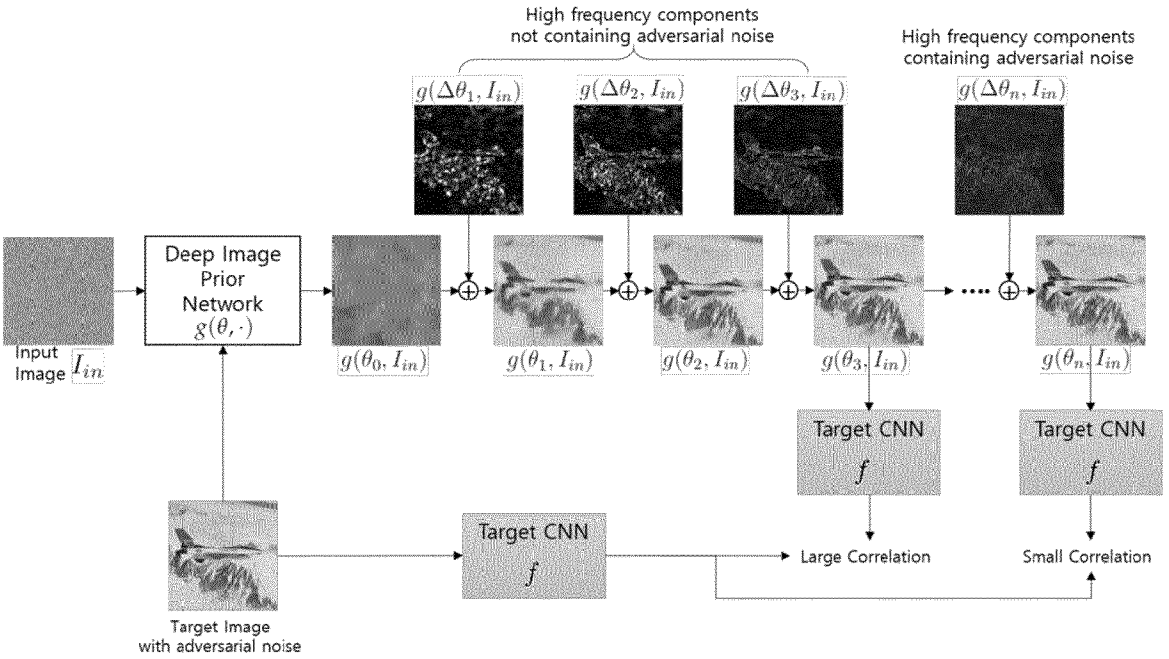


(a) Conventional Deep Image Prior

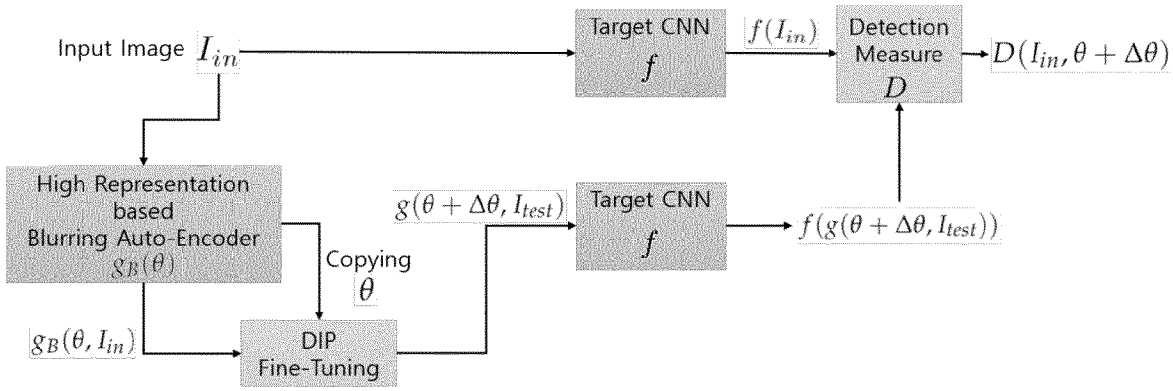


(b) Deep Image Prior with Blurring Network Parameter Initialization

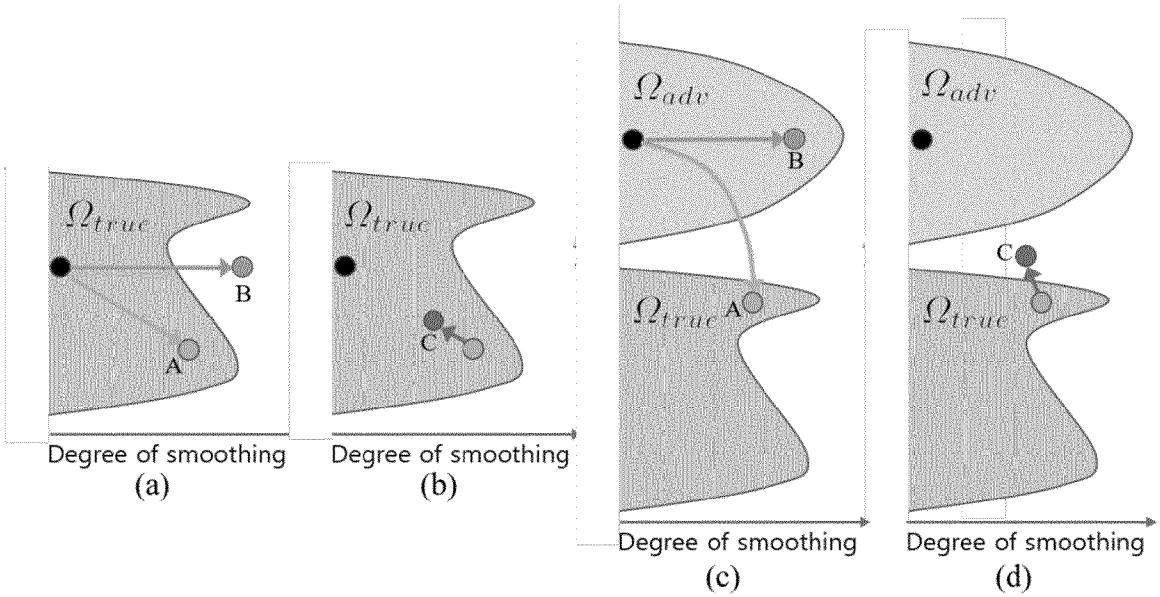
[도3]


High frequency components not containing adversarial noise
High frequency components containing adversarial noise
$g(\Delta\theta_1, I_{in})$
$g(\Delta\theta_2, I_{in})$
$g(\Delta\theta_3, I_{in})$
$g(\Delta\theta_n, I_{in})$
Input Image I_{in}
Deep Image Prior Network $g(\theta, \cdot)$
$g(\theta_0, I_{in})$
$g(\theta_1, I_{in})$
$g(\theta_2, I_{in})$
$g(\theta_3, I_{in})$
$g(\theta_n, I_{in})$
Target CNN f
Target CNN f
Target CNN f
Large Correlation
Small Correlation
Target Image with adversarial noise

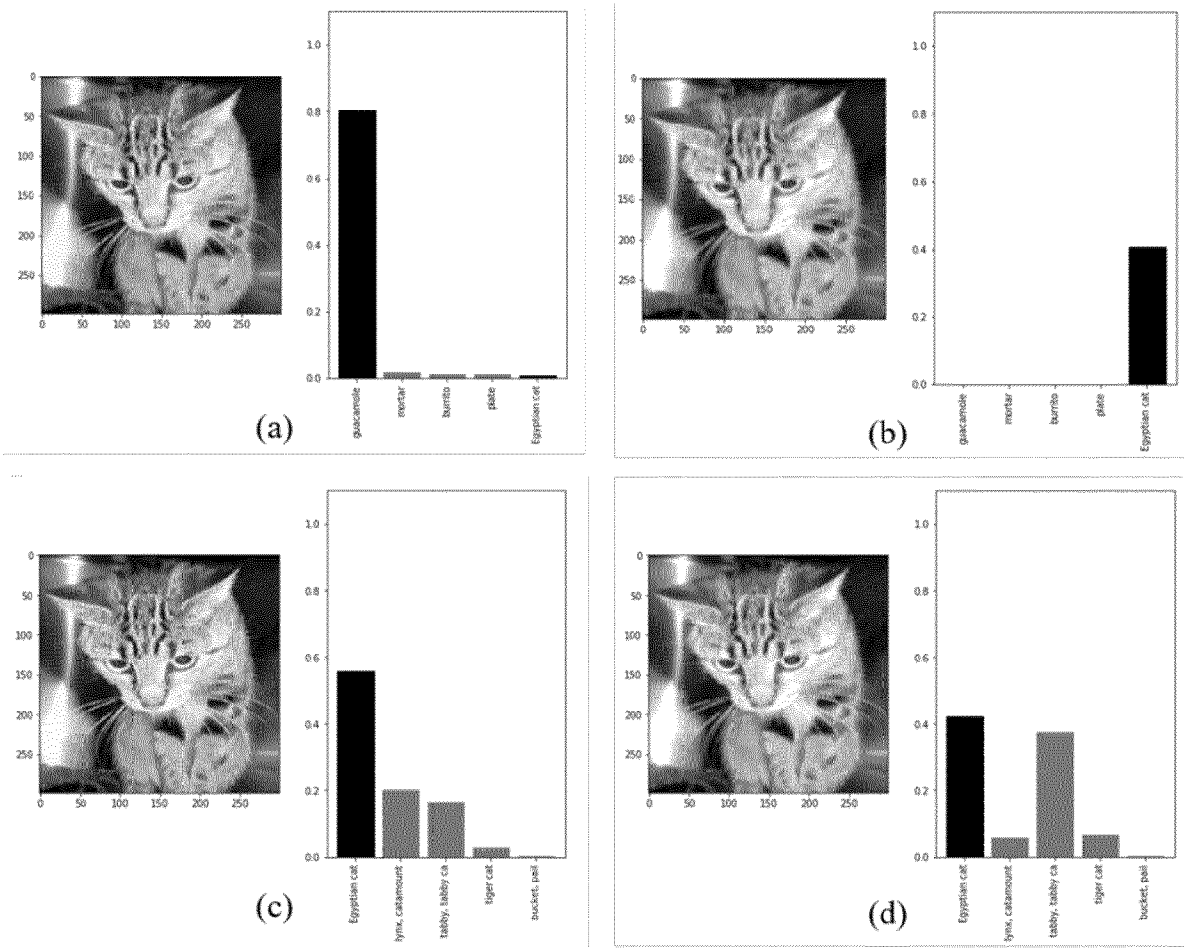
[도4]


Input Image I_{in}
Target CNN f
$f(I_{in})$
Detection Measure D
$D(I_{in}, \theta + \Delta\theta)$
High Representation based Blurring Auto-Encoder $g_B(\theta)$
Copying θ
$g(\theta + \Delta\theta, I_{test})$
Target CNN f
$f(g(\theta + \Delta\theta, I_{test}))$
$g_B(\theta, I_{in})$
DIP Fine-Tuning

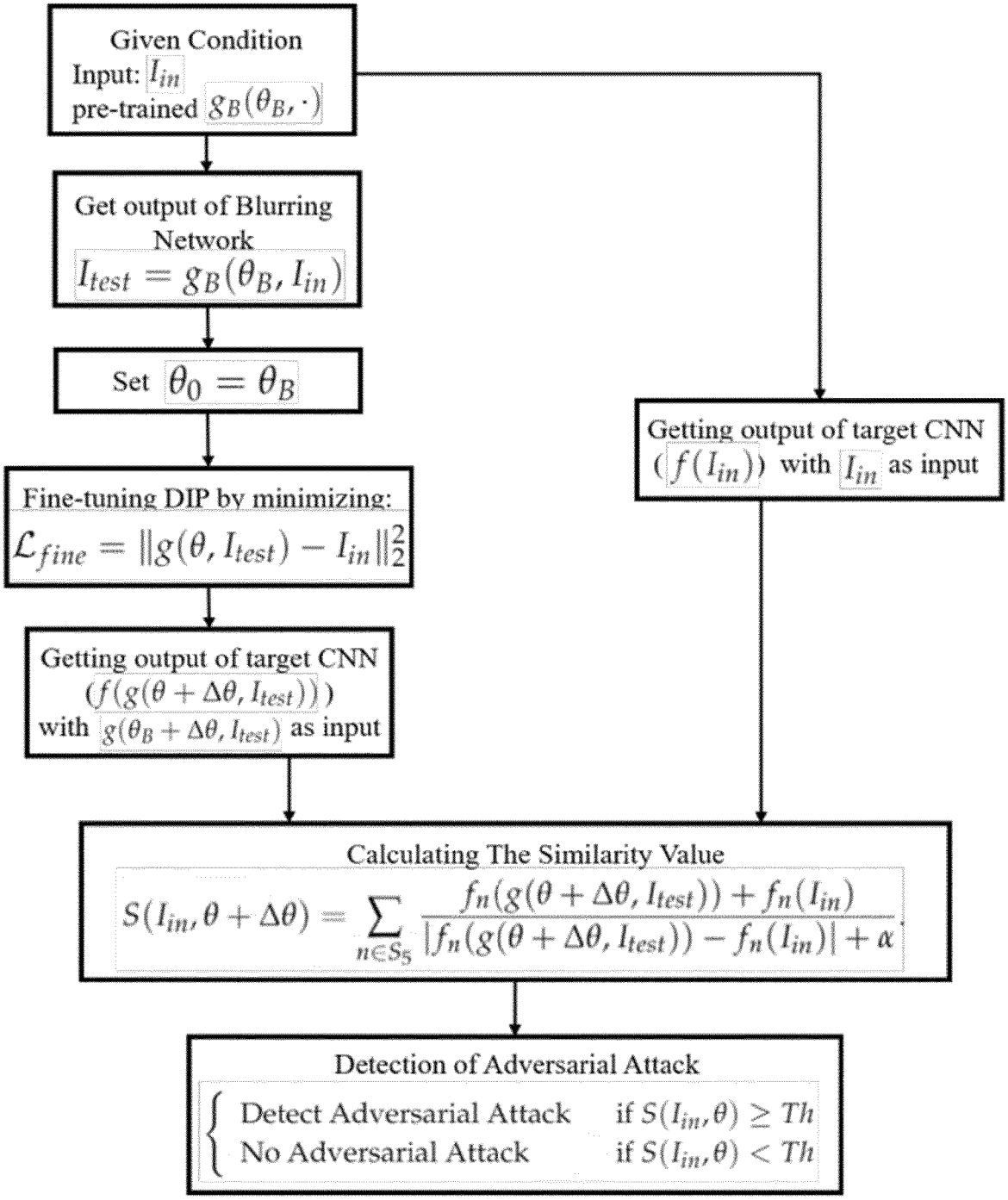
[도5]


Ω_{true}
A
B
Degree of smoothing
(a)
Ω_{true}
C
Degree of smoothing
(b)
Ω_{adv}
B
Ω_{true}
A
Degree of smoothing
(c)
Ω_{adv}
C
Ω_{true}
Degree of smoothing
(d)

[도6]


(a)
guacamole
mortar
burrito
plate
Egyptian cat
(b)
guacamole
mortar
burrito
plate
Egyptian cat
(c)
Egyptian cat
lynx, catamount
tabby, tabby ca
tiger cat
bucket, pail
(d)
Egyptian cat
lynx, catamount
tabby, tabby ca
tiger cat
bucket, pail

[도7]


Given Condition
Input: I_{in}
pre-trained $g_B(\theta_B, \cdot)$
Get output of Blurring Network
$I_{test} = g_B(\theta_B, I_{in})$
Set $\theta_0 = \theta_B$
Fine-tuning DIP by minimizing:
$\mathcal{L}_{fine} = \|g(\theta, I_{test}) - I_{in}\|_2^2$
Getting output of target CNN $(f(g(\theta + \Delta\theta, I_{test})))$ with $g(\theta_B + \Delta\theta, I_{test})$ as input
Getting output of target CNN $(f(I_{in}))$ with I_{in} as input
Calculating The Similarity Value
$S(I_{in}, \theta + \Delta\theta) = \sum_{n \in S_5} \frac{f_n(g(\theta + \Delta\theta, I_{test})) + f_n(I_{in})}{|f_n(g(\theta + \Delta\theta, I_{test})) - f_n(I_{in})| + \alpha}$
Detection of Adversarial Attack
Detect Adversarial Attack if $S(I_{in}, \theta) \geq Th$
No Adversarial Attack if $S(I_{in}, \theta) < Th$

INTERNATIONAL SEARCH REPORT

International application No.
PCT/KR2021/008127

A. CLASSIFICATION OF SUBJECT MATTER

G06F 21/55(2013.01)i; **G06N 3/08**(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F 21/55(2013.01); G06N 3/04(2006.01); G06N 3/08(2006.01); G06T 1/20(2006.01)

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Korean utility models and applications for utility models: IPC as above
Japanese utility models and applications for utility models: IPC as above

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

eKOMPASS (KIPO internal) & keywords: 적대적(adversarial), 잡음(noise), 공격(attack), Deep Image Prior(DIP), 블러링(blurring)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	SUTANTO, Richard Evan et al. Real-Time Adversarial Attack Detection with Deep Image Prior Initialized as a High-Level Representation Based Blurring Network. Electronics 2021. Volume 10, Issue 1, 52, pp. 1-17, 30 December 2020 [Retrieved on 03 May 2022]. Retrieved from <URL: https://www.mdpi.com/2079-9292/10/1/52>. See pages 2 and 4-5.	1-3
X	SHI, Yu et al. Unsupervised Adversarial Defense through Tandem Deep Image Priors. Electronics 2020. Volume 9, Issue 11, 1957, pp. 1-13, 19 November 2020 [Retrieved on 03 May 2022]. Retrieved from < URL: https://www.mdpi.com/2079-9292/9/11/1957>. See page 2.	1
A	KR 10-2020-0080424 A (NAVER CORPORATION) 07 July 2020 (2020-07-07) See paragraphs [0038]-[0073]; claim 1; and figures 2-5.	1-3
A	US 2021-0192339 A1 (DONGSEO UNIVERSITY HEADQUARTERS) 24 June 2021 (2021-06-24) See claims 1-4; and figures 1-2.	1-3

☑ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"D" document cited by the applicant in the international application
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
09 May 2022	**10 May 2022**

Name and mailing address of the ISA/KR	Authorized officer
Korean Intellectual Property Office **Government Complex-Daejeon Building 4, 189 Cheongsa-ro, Seo-gu, Daejeon 35208**	
Facsimile No. **+82-42-481-8578**	Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/KR2021/008127

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	KR 10-2191722 B1 (INDUSTRY ACADEMY COOPERATION FOUNDATION OF SEJONG UNIVERSITY) 16 December 2020 (2020-12-16) See claims 1-3; and figures 1-3.	1-3

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/KR2021/008127

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
KR 10-2020-0080424 A	07 July 2020	None	
US 2021-0192339 A1	24 June 2021	KR 10-2021-0081769 A	02 July 2021
		KR 10-2304661 B1	23 September 2021
KR 10-2191722 B1	16 December 2020	US 2022-0019675 A1	20 January 2022

국 제 조 사 보 고 서

국제출원번호
PCT/KR2021/008127

A. 발명이 속하는 기술분류(국제특허분류(IPC))

G06F 21/55(2013.01)i; **G06N 3/08**(2006.01)i

B. 조사된 분야

조사된 최소문헌(국제특허분류를 기재)

G06F 21/55(2013.01); G06N 3/04(2006.01); G06N 3/08(2006.01); G06T 1/20(2006.01)

조사된 기술분야에 속하는 최소문헌 이외의 문헌

한국등록실용신안공보 및 한국공개실용신안공보: 조사된 최소문헌란에 기재된 IPC
일본등록실용신안공보 및 일본공개실용신안공보: 조사된 최소문헌란에 기재된 IPC

국제조사에 이용된 전산 데이터베이스(데이터베이스의 명칭 및 검색어(해당하는 경우))

eKOMPASS(특허청 내부 검색시스템) & 키워드: 적대적(adversarial), 잡음(noise), 공격(attack), Deep Image Prior(DIP), 블러링(blurring)

C. 관련 문헌

카테고리*	인용문헌명 및 관련 구절(해당하는 경우)의 기재	관련 청구항
X	RICHARD EVAN SUTANTO 등, `Real-Time Adversarial Attack Detection with Deep Image Prior Initialized as a High-Level Representation Based Blurring Network`, Electronics 2021, Volume 10, Issue 1, 52, pp. 1-17, 2020.12.30 [검색일: 2022.05.03]. 출처: <URL: https://www.mdpi.com/2079-9292/10/1/52> 페이지 2, 4-5	1-3
X	YU SHI 등, `Unsupervised Adversarial Defense through Tandem Deep Image Priors`, Electronics 2020, Volume 9, Issue 11, 1957, pp. 1-13, 2020.11.19 [검색일: 2022.05.03]. 출처: <URL: https://www.mdpi.com/2079-9292/9/11/1957> 페이지 2	1
A	KR 10-2020-0080424 A (네이버 주식회사) 2020.07.07 단락 [0038]-[0073]; 청구항 1; 및 도면 2-5	1-3
A	US 2021-0192339 A1 (DONGSEO UNIVERSITY HEADQUARTERS) 2021.06.24 청구항 1-4; 및 도면 1-2	1-3
A	KR 10-2191722 B1 (세종대학교산학협력단) 2020.12.16 청구항 1-3; 및 도면 1-3	1-3

☐ 추가 문헌이 C(계속)에 기재되어 있습니다.

☑ 대응특허에 관한 별지를 참조하십시오.

* 인용된 문헌의 특별 카테고리:

"A" 특별히 관련이 없는 것으로 보이는 일반적인 기술수준을 정의한 문헌

"D" 본 국제출원에서 출원인이 인용한 문헌

"E" 국제출원일보다 빠른 출원일 또는 우선일을 가지나 국제출원일 이후에 공개된 선출원 또는 특허 문헌

"L" 우선권 주장에 의문을 제기하는 문헌 또는 다른 인용문헌의 공개일 또는 다른 특별한 이유(이유를 명시)를 밝히기 위하여 인용된 문헌

"O" 구두 개시, 사용, 전시 또는 기타 수단을 언급하고 있는 문헌

"P" 우선일 이후에 공개되었으나 국제출원일 이전에 공개된 문헌

"T" 국제출원일 또는 우선일 후에 공개된 문헌으로, 출원과 상충하지 않으며 발명의 기초가 되는 원리나 이론을 이해하기 위해 인용된 문헌

"X" 특별한 관련이 있는 문헌. 해당 문헌 하나만으로 청구된 발명의 신규성 또는 진보성이 없는 것으로 본다.

"Y" 특별한 관련이 있는 문헌. 해당 문헌이 하나 이상의 다른 문헌과 조합하는 경우로 그 조합이 당업자에게 자명한 경우 청구된 발명은 진보성이 없는 것으로 본다.

"&" 동일한 대응특허문헌에 속하는 문헌

국제조사의 실제 완료일	국제조사보고서 발송일
2022년05월09일(09.05.2022)	**2022년05월10일(10.05.2022)**

ISA/KR의 명칭 및 우편주소

대한민국 특허청
(35208) 대전광역시 서구 청사로 189, 4동 (둔산동, 정부대전청사)

팩스 번호 **+82-42-481-8578**

심사관

양정록

전화번호 **+82-42-481-5709**

국 제 조 사 보 고 서
대응특허에 관한 정보

국제출원번호
PCT/KR2021/008127

국제조사보고서에서 인용된 특허문헌	공개일	대응특허문헌	공개일
KR 10-2020-0080424 A	2020/07/07	없음	
US 2021-0192339 A1	2021/06/24	KR 10-2021-0081769 A	2021/07/02
		KR 10-2304661 B1	2021/09/23
KR 10-2191722 B1	2020/12/16	US 2022-0019675 A1	2022/01/20