(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2023년11월06일
(11) 등록번호 10-2598816
(24) 등록일자 2023년11월01일

(51) 국제특허분류(Int. Cl.)
G06F 18/00 (2023.01) *G06N 3/08* (2023.01)
(52) CPC특허분류
G06V 20/46 (2022.01)
G06N 3/08 (2023.01)
(21) 출원번호 10-2021-0141638
(22) 출원일자 2021년10월22일
심사청구일자 2021년10월22일
(65) 공개번호 10-2023-0057647
(43) 공개일자 2023년05월02일
(56) 선행기술조사문헌
D. Tran et al., 'Learning Spatiotemporal Features with 3D Convolutional Networks,' arXiv:1412.0767v4 [cs.CV] (2015.10.07.) 1부.*
I. Goodfellow et al., 'Explaining and harnessing Adversarial Examples,' arXiv:1412.6572v3 [stat.ML] (2015.03.20.) 1부.*
머시러닝 보안 취약점! 적대적 공격의 4가지 유형,' LG CNS 블로그, 2020.02.13. (online: blog.lgcns.com/2191) 1부.*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
연세대학교 산학협력단
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
(72) 발명자
이종석
인천광역시 연수구 송도과학로 85 연세대학교 국제캠퍼스
황재희
인천광역시 연수구 송도과학로27번길 55 롯데캐슬 캠퍼스타운 202동 943호
(뒷면에 계속)
(74) 대리인
정부연

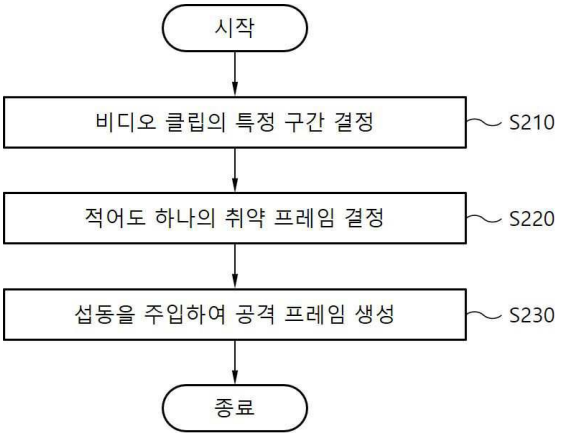
전체 청구항 수 : 총 11 항 심사관 : 황승희

(54) 발명의 명칭 **행동인식을 위한 딥러닝 모델에 대한 인지 최적화 적대적 공격 장치 및 방법**

(57) 요 약

본 발명은 행동인식을 위한 인지 최적화 적대적 공격 장치 및 방법에 관한 것으로, 상기 장치는 대상 행동인식 모델을 통해 정상적으로 인식되는 오리지널 비디오 클립의 특정 구간을 결정하는 정상인식 구간 결정부; 상기 특정 구간에서 원본 프레임과의 차이를 특정 기준 이하로 유지시키는 섭동(perturbation)을 통해 상기 대상 행동인식 모델의 오인식을 발생시키는 적어도 하나의 취약 프레임을 검출하는 취약 프레임 검출부; 및 상기 적어도 하나의 취약 프레임에만 상기 섭동을 주입하는 화이트박스 시나리오에 따라 상기 적어도 하나의 취약 프레임에 대응되는 공격 프레임을 생성하는 공격 프레임 생성부;를 포함한다.

대 표 도 - 도2



(52) CPC특허분류
G06V 40/20 (2022.01)
(72) 발명자
최준호
경기도 의왕시 포일세거리로 10 포일프라임타워 618

김준혁
인천광역시 연수구 송도과학로27번길 55 롯데캐슬 캠퍼스타운 201동 1102호

이 발명을 지원한 국가연구개발사업

과제고유번호	1711126082
과제번호	2020-0-01361-002
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신방송혁신인재양성(R&D)
연구과제명	인공지능대학원지원(연세대학교)
기 여 율	1/1
과제수행기관명	연세대학교 산학협력단
연구기간	2021.01.01 ~ 2021.12.31

공지예외적용 : 있음

명 세 서

청구범위

청구항 1

대상 행동인식 모델을 통해 정상적으로 인식되는 오리지널 비디오 클립의 제1 재생시점부터 제2 재생시점까지의 연속적인 프레임들을 특정 구간으로 결정하는 정상인식 구간 결정부;

상기 특정 구간의 연속적인 프레임들 중 어느 하나에 원본 프레임과의 차이를 특정 기준 이하로 유지시키는 섭동(perturbation)을 주입하여 상기 대상 행동인식 모델의 오인식 가능성이 가장 높은 적어도 하나의 취약 프레임을 검출하는 취약 프레임 검출부; 및

상기 적어도 하나의 취약 프레임에만 상기 섭동을 주입하는 화이트박스 시나리오에 따라 상기 적어도 하나의 취약 프레임에 대응되는 공격 프레임을 생성하는 공격 프레임 생성부;를 포함하되,

상기 취약 프레임 검출부는 상기 섭동이 주입된 프레임을 상기 대상 행동인식 모델에 입력하여 상기 섭동에 관한 해당 프레임의 속임률(Fooling rate)을 산출하는 것을 특징으로 하는 행동인식을 위한 인지 최적화 적대적 공격 장치.

청구항 2

삭제

청구항 3

제1항에 있어서, 상기 취약 프레임 검출부는

상기 연속적인 프레임들 중 어느 하나에 상기 섭동으로서 균일 랜덤 노이즈(uniform random noise)를 주입하는 것을 특징으로 하는 행동인식을 위한 인지 최적화 적대적 공격 장치.

청구항 4

제1항에 있어서, 상기 취약 프레임 검출부는

I-FGSM 알고리즘을 적용하여 상기 연속적인 프레임들 중 어느 하나에 상기 섭동을 주입하는 것을 특징으로 하는 행동인식을 위한 인지 최적화 적대적 공격 장치.

청구항 5

제4항에 있어서, 상기 취약 프레임 검출부는

상기 I-FGSM 알고리즘의 반복마다 다음의 수학식 1을 통해 상기 연속적인 프레임들에서 i번째 프레임을 검출하는 것을 특징으로 하는 행동인식을 위한 인지 최적화 적대적 공격 장치.

[수학식 1]

$$\tilde{X}^{n+1}(i) = Clip_{0,255}\left(X^n(i) + \frac{\epsilon}{N} sgn\left(\nabla_{X^n(i)} J(X^n, y)\right)\right)$$

$$X^{n+1}(i) = Clip_{-\epsilon,\epsilon}\left(\tilde{X}^{n+1}(i) - X^0(i)\right) + X^0(i)$$

(여기에서, ϵ는 추가될 섭동량이고, sgn(·)은 부호 함수이고, $\nabla_{X^n(i)} J(X^n, y)$는 손실 함수에 대한 대상 프레

임의 그라디언트이다.)

청구항 6

제4항에 있어서, 상기 취약 프레임 검출부는

복수의 오리지널 비디오 클립들에 대한 보편적 섭동(universial perturbation)을 통해 상기 적어도 하나의 취약 프레임을 검출하는 것을 특징으로 하는 행동인식을 위한 인지 최적화 적대적 공격 장치.

청구항 7

삭제

청구항 8

제1항에 있어서, 상기 취약 프레임 검출부는

상기 연속적인 프레임들 각각에 대해 상기 속임률을 산출하고 상기 속임률을 기초로 상기 적어도 하나의 취약 프레임을 결정하는 것을 특징으로 하는 행동인식을 위한 인지 최적화 적대적 공격 장치.

청구항 9

제8항에 있어서, 상기 공격 프레임 생성부는

상기 적어도 하나의 취약 프레임 중에서 상기 속임률이 가장 높은 프레임에 대해 단일 프레임 공격을 통해 상기 공격 프레임을 생성하는 것을 특징으로 하는 행동인식을 위한 인지 최적화 적대적 공격 장치.

청구항 10

제1항에 있어서,

상기 오리지널 비디오 클립의 특정 구간에 상기 공격 프레임을 대체 삽입하여 상기 대상 행동인식 모델에 대한 공격 성능을 평가하는 공격 성능 분석부;를 더 포함하는 것을 특징으로 하는 행동인식을 위한 인지 최적화 적대적 공격 장치.

청구항 11

제10항에 있어서, 상기 공격 성능 분석부는

속임률(fooling rate)과 섭동의 눈에 띄지 않는 정도(inconspicuousness)를 기초로 상기 공격 성능을 평가하는 것을 특징으로 하는 행동인식을 위한 인지 최적화 적대적 공격 장치.

청구항 12

적대적 공격 장치에서 수행되는 적대적 공격 방법에 있어서,

정상인식 구간 결정부를 통해, 대상 행동인식 모델을 통해 정상적으로 인식되는 오리지널 비디오 클립의 제1 재생시점부터 제2 재생시점까지의 연속적인 프레임들을 특정 구간으로 결정하는 정상인식 구간 결정단계;

취약 프레임 검출부를 통해, 상기 특정 구간의 연속적인 프레임들 중 어느 하나에 원본 프레임과의 차이를 특정 기준 이하로 유지시키는 섭동(perturbation)을 주입하여 상기 대상 행동인식 모델의 오인식 가능성이 가장 높은 적어도 하나의 취약 프레임을 검출하는 취약 프레임 검출단계; 및

공격 프레임 생성부를 통해, 상기 적어도 하나의 취약 프레임에만 상기 섭동을 주입하는 화이트박스 시나리오에 따라 상기 적어도 하나의 취약 프레임에 대응되는 공격 프레임을 생성하는 공격 프레임 생성단계;를 포함하되,

상기 취약 프레임 검출단계는 상기 섭동이 주입된 프레임을 상기 대상 행동인식 모델에 입력하여 상기 섭동에 관한 해당 프레임의 속임률(Fooling rate)을 산출하는 단계를 포함하는 것을 특징으로 하는 행동인식을 위한 인지 최적화 적대적 공격 방법.

청구항 13

제12항에 있어서,

공격 성능 분석부를 통해, 상기 오리지널 비디오 클립의 특정 구간에 상기 공격 프레임을 대체 삽입하여 상기 대상 행동인식 모델에 대한 공격 성능을 평가하는 공격 성능 분석단계;를 더 포함하는 것을 특징으로 하는 행동인식을 위한 인지 최적화 적대적 공격 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 적대적 공격 기술에 관한 것으로, 보다 상세하게는 비디오 클립의 프레임에 섭동을 주입하여 딥 러닝 기반 행동인식 모델의 구조적 취약성을 공격할 수 있는 인지 최적화 적대적 공격 장치 및 방법에 관한 것이다.

배 경 기 술

[0003] 적대적 공격(adversarial attack)이라고 불리는, 입력 데이터에 작은 섭동(small perturbation)을 주입하면 심층 신경망(deep neural network)의 성능이 크게 저하될 수 있다. 이러한 공격은 딥러닝(deep learning) 기반 응용들(applications)에 대한 보안 문제를 제기하기 때문에 많은 연구자들에 의해 다양한 딥 모델들, 특히 이미지 분류 모델(image classification model)들에 대한 적대적 공격의 영향이 연구되어 왔다.

[0004] 영상을 이용한 인간 행동인식은 풍부한 연산 자원과 데이터를 기반으로 하는 심층 네트워크 기반 알고리즘의 개발로 인해 최근 몇 년 동안 광범위하게 연구되고 있다. 네트워크 설계 관점에서 행동인식의 주요 연구 주제는 비디오 클립에 있는 시간 정보를 모델링하는 방법일 수 있다. 이와 관련하여 LSTM(장단기기억) 모듈이나 광 흐름(optical flow)을 활용하는 등 다양한 시도가 있었지만 최근에는 3D CNN(Convolutional Neural Network) 기반의 행동인식 모델이 널리 보급되고 있다. 사용된 3D CNN 기반 행동인식 모델의 성능과 효율성을 향상시키기 위해 프레임 선택(frame selection) 및 컨볼루션 연산(convolutional operation)과 같은 시간 차원의 다양한 메커니즘이 제안되고 있다.

[0005] 또한, 많은 연구자들에 의해 입력 데이터에 눈에 띄지 않는 교란을 추가하여 대상 모델이 잘못된 출력을 생성하도록 유도하는 소위 적대적 공격에 대한 딥러닝 기반 알고리즘의 취약성(vulnerability)이 발견되고 있다.

선행기술문헌

특허문헌

[0007] (특허문헌 0001) 한국공개특허 제10-2019-0061446호 (2019.06.05)

발명의 내용

해결하려는 과제

[0008] 본 발명의 일 실시예는 비디오 클립의 단일 프레임에 눈에 띄지 않는 섭동을 주입하는 단일 프레임 공격을 통해 딥 러닝 기반 행동인식 모델의 구조적 취약성을 공격할 수 있는 인지 최적화 적대적 공격 장치 및 방법을 제공하고자 한다.

과제의 해결 수단

[0010] 실시예들 중에서, 행동인식을 위한 인지 최적화 적대적 공격 장치는 대상 행동인식 모델을 통해 정상적으로 인식되는 오리지널 비디오 클립의 특정 구간을 결정하는 정상인식 구간 결정부; 상기 특정 구간에서 원본 프레임과의 차이를 특정 기준 이하로 유지시키는 섭동(perturbation)을 통해 상기 대상 행동인식 모델의 오인식을 발생시키는 적어도 하나의 취약 프레임을 검출하는 취약 프레임 검출부; 및 상기 적어도 하나의 취약 프레임에만 상기 섭동을 주입하는 화이트박스 시나리오에 따라 상기 적어도 하나의 취약 프레임에 대응되는 공격 프레임을 생성하는 공격 프레임 생성부;를 포함한다.

[0011] 상기 정상인식 구간 결정부는 상기 오리지널 비디오 클립의 제1 재생시점부터 제2 재생시점까지의 연속적인 프레임들로 상기 특정 구간을 구성할 수 있다.

[0012] 상기 취약 프레임 검출부는 상기 연속적인 프레임들 중 어느 하나에 상기 섭동으로서 균일 랜덤 노이즈(uniform random noise)를 주입할 수 있다.

[0013] 상기 취약 프레임 검출부는 I-FGSM 알고리즘을 적용하여 상기 연속적인 프레임들 중 어느 하나에 상기 섭동을 주입할 수 있다.

[0014] 상기 취약 프레임 검출부는 상기 I-FGSM 알고리즘의 반복마다 다음의 수학식 1을 통해 상기 연속적인 프레임들에서 i번째 프레임을 검출할 수 있다.

[0015] [수학식 1]

[0016] $$\tilde{X}^{n+1}(i) = Clip_{0,255}\left(X^n(i) + \frac{\epsilon}{N} sgn\left(\nabla_{X^n(i)} J(X^n, y)\right)\right)$$

[0017] $$X^{n+1}(i) = Clip_{-\epsilon,\epsilon}\left(\tilde{X}^{n+1}(i) - X^0(i)\right) + X^0(i)$$

[0018] (여기에서, ϵ는 추가될 섭동량이고, sgn(·)은 부호 함수이고, $\nabla_{X^n(i)} J(X^n, y)$는 손실 함수에 대한 대상 프레임의 그라디언트이다.)

[0019] 상기 취약 프레임 검출부는 복수의 오리지널 비디오 클립들에 대한 보편적 섭동(universial perturbation)을 통해 상기 적어도 하나의 취약 프레임을 검출할 수 있다.

[0020] 상기 취약 프레임 검출부는 상기 섭동이 주입된 프레임을 상기 대상 행동인식 모델에 입력하여 상기 섭동에 관한 해당 프레임의 속임률(Fooling rate)을 산출할 수 있다.

[0021] 상기 취약 프레임 검출부는 상기 연속적인 프레임들 각각에 대해 상기 속임률을 산출하고 상기 속임률을 기초로 상기 적어도 하나의 취약 프레임을 결정할 수 있다.

[0022] 상기 공격 프레임 생성부는 상기 적어도 하나의 취약 프레임 중에서 상기 속임률이 가장 높은 프레임에 대해 단일 프레임 공격을 통해 상기 공격 프레임을 생성할 수 있다.

[0023] 상기 장치는 상기 오리지널 비디오 클립의 특정 구간에 상기 공격 프레임을 대체 삽입하여 상기 대상 행동인식 모델에 대한 공격 성능을 평가하는 공격 성능 분석부;를 더 포함할 수 있다.

[0024] 상기 공격 성능 분석부는 속임률(fooling rate)과 섭동의 눈에 띄지 않는 정도(inconspicuousness)를 기초로 상기 공격 성능을 평가할 수 있다.

[0025] 실시예들 중에서, 행동인식을 위한 인지 최적화 적대적 공격 방법은 대상 행동인식 모델을 통해 정상적으로 인식되는 오리지널 비디오 클립의 특정 구간을 결정하는 정상인식 구간 결정단계; 상기 특정 구간에서 원본 프레임과의 차이를 특정 기준 이하로 유지시키는 섭동(perturbation)을 통해 상기 대상 행동인식 모델의 오인식을 발생시키는 적어도 하나의 취약 프레임을 검출하는 취약 프레임 검출단계; 및 상기 적어도 하나의 취약 프레임에만 상기 섭동을 주입하는 화이트박스 시나리오에 따라 상기 적어도 하나의 취약 프레임에 대응되는 공격 프레임을 생성하는 공격 프레임 생성단계;를 포함한다.

[0026] 상기 방법은 상기 오리지널 비디오 클립의 특정 구간에 상기 공격 프레임을 대체 삽입하여 상기 대상 행동인식 모델에 대한 공격 성능을 평가하는 공격 성능 분석단계;를 더 포함할 수 있다.

발명의 효과

[0028] 개시된 기술은 다음의 효과를 가질 수 있다. 다만, 특정 실시예가 다음의 효과를 전부 포함하여야 한다거나 다음의 효과만을 포함하여야 한다는 의미는 아니므로, 개시된 기술의 권리범위는 이에 의하여 제한되는 것으로 이해되어서는 아니 될 것이다.

[0029] 본 발명에 따른 인지 최적화 적대적 공격 장치 및 방법은 비디오 클립의 단일 프레임에 눈에 띄지 않는 섭동을 주입하는 단일 프레임 공격을 통해 딥 러닝 기반 행동인식 모델의 구조적 취약성을 공격할 수 있다.

도면의 간단한 설명

[0031] 도 1은 본 발명에 따른 적대적 공격 장치의 물리적 구성을 설명하는 도면이다.

도 2는 본 발명에 따른 행동인식을 위한 인지 최적화 적대적 공격 방법을 설명하는 순서도이다.

도 3은 I-FGSM에 의한 단일 프레임의 섭동에 관한 속임률을 설명하는 도면이다.

도 4는 균일 랜덤 노이즈에 의한 단일 프레임의 섭동에 관한 속임률을 설명하는 도면이다.

도 5는 본 발명에 따른 행동인식 모델의 구조적 취약성을 설명하는 도면이다.

도 6은 행동인식 모델에서 서로 다른 프레임 간의 전이성을 설명하는 도면이다.

도 7은 행동인식 모델들 간의 단일 프레임 공격의 속임률을 비교 설명하는 도면이다.

도 8은 I3D에서 섭동의 양에 따른 가시성과 탐지율을 설명하는 도면이다.

도 9는 탐지율에 관한 실험 결과를 설명하는 도면이다.

도 10은 범용 단일 프레임 공격에 대한 실험 결과를 설명하는 도면이다.

도 11은 시간 불변의 범용 공격의 속임률을 설명하는 도면이다.

도 12는 본 발명에 따른 적대적 공격 장치의 시스템 구성을 설명하는 도면이다.

도 13은 본 발명에 따른 적대적 공격 시스템을 설명하는 도면이다.

발명을 실시하기 위한 구체적인 내용

[0032] 본 발명에 관한 설명은 구조적 내지 기능적 설명을 위한 실시예에 불과하므로, 본 발명의 권리범위는 본문에 설명된 실시예에 의하여 제한되는 것으로 해석되어서는 아니 된다. 즉, 실시예는 다양한 변경이 가능하고 여러 가지 형태를 가질 수 있으므로 본 발명의 권리범위는 기술적 사상을 실현할 수 있는 균등물들을 포함하는 것으로 이해되어야 한다. 또한, 본 발명에서 제시된 목적 또는 효과는 특정 실시예가 이를 전부 포함하여야 한다거나 그러한 효과만을 포함하여야 한다는 의미는 아니므로, 본 발명의 권리범위는 이에 의하여 제한되는 것으로 이해되어서는 아니 될 것이다.

[0033] 한편, 본 출원에서 서술되는 용어의 의미는 다음과 같이 이해되어야 할 것이다.

[0034] "제1", "제2" 등의 용어는 하나의 구성요소를 다른 구성요소로부터 구별하기 위한 것으로, 이들 용어들에 의해 권리범위가 한정되어서는 아니 된다. 예를 들어, 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다.

[0035] 어떤 구성요소가 다른 구성요소에 "연결되어"있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결될 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성요소가 다른 구성요소에 "직접 연결되어"있다고 언급된 때에는 중간에 다른 구성요소가 존재하지 않는 것으로 이해되어야 할 것이다. 한편, 구성요소들 간의 관계를 설명하는 다른 표현들, 즉 "~사이에"와 "바로 ~사이에" 또는 "~에 이웃하는"과 "~에 직접 이웃하는" 등도 마찬가지로 해석되어야 한다.

[0036] 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한 복수의 표현을 포함하는 것으로 이해되어야 하고, "포함하다"또는 "가지다" 등의 용어는 실시된 특징, 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함을 지정하려는 것이며, 하나 또는 그 이상의 다른 특징이나 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0037] 각 단계들에 있어 식별부호(예를 들어, a, b, c 등)는 설명의 편의를 위하여 사용되는 것으로 식별부호는 각 단계들의 순서를 설명하는 것이 아니며, 각 단계들은 문맥상 명백하게 특정 순서를 기재하지 않는 이상 명기된 순서와 다르게 일어날 수 있다. 즉, 각 단계들은 명기된 순서와 동일하게 일어날 수도 있고 실질적으로 동시에 수행될 수도 있으며 반대의 순서대로 수행될 수도 있다.

[0038] 본 발명은 컴퓨터가 읽을 수 있는 기록매체에 컴퓨터가 읽을 수 있는 코드로서 구현될 수 있고, 컴퓨터가 읽을 수 있는 기록 매체는 컴퓨터 시스템에 의하여 읽혀질 수 있는 데이터가 저장되는 모든 종류의 기록 장치를 포함한다. 컴퓨터가 읽을 수 있는 기록 매체의 예로는 ROM, RAM, CD-ROM, 자기 테이프, 플로피 디스크, 광 데이터 저장 장치 등이 있다. 또한, 컴퓨터가 읽을 수 있는 기록 매체는 네트워크로 연결된 컴퓨터 시스템에 분산되어, 분산 방식으로 컴퓨터가 읽을 수 있는 코드가 저장되고 실행될 수 있다.

[0039] 여기서 사용되는 모든 용어들은 다르게 정의되지 않는 한, 본 발명이 속하는 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 일반적으로 사용되는 사전에 정의되어 있는 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한 이상적이거나 과도하게 형식적인 의미를 지니는 것으로 해석될 수 없다.

[0041] 최근 심층 신경망(deep neural network)의 발달과 함께 행동인식(action recognition)의 성능이 크게 향상되고 있다. 초기의 CNN+LSTM 구조는 이미지 관련 작업에서 널리 사용되는 2차원 컨볼루션 레이어와 시퀀스 데이터를 대상으로 하는 LSTM 모델을 통합하여 높은 성능을 달성하였다. 한편, 공간적 차원과 시간적 차원 모두에서 특징들을 활용하는 3차원(3D) 컨볼루션 레이어도 제안되었다. 비디오를 처리하는 또 다른 접근 방식은 두 개의 CNN을 동시에 사용하는 것일 수 있다(2-스트림 네트워크로 알려짐). 해당 방식은 원본 RGB 프레임(공간적 특징을 이용하기 위해)과 광학적 흐름(시간적 특징을 이용하기 위해)을 각각 처리할 수 있다. 이 두 가지 접근 방식은 때때로 행동인식의 성능을 더욱 향상시키기 위해 결합될 수 있다.

[0042] 최근 몇 년 동안에는 보다 발전된 딥 행동인식 모델(deep action recognition model)이 개발되어 왔다. 널리 사용되는 접근 방식 중 하나는 팽창된 3차원(I3D) 네트워크이며, 이는 컨볼루션 레이어의 2차원 커널을 3차원으로 팽창시킨다. 2-스트림 접근(two-stream approach) 방식은 각각의 스트림 네트워크에서 서로 다른 시간 해상도(즉, 프레임 속도)를 갖는 비디오 데이터를 입력으로 사용하는 SlowFast라는 방법으로 확장되었다. 행동인식에 대한 최근 연구의 또 다른 경향은 계산 복잡성(computational complexity)을 줄이기 위해 커널 인수분해(kernel factorization)(예: 상호 작용 감소 채널 분리 네트워크(ir-CSN))를 사용하는 것일 수 있다.

[0043] 딥러닝 기반 이미지 분류 모델은 화이트 박스 시나리오(white-box scenario) 하에서 적대적 공격(adversarial attack)에 매우 취약할 수 있다. 주어진 모델의 분류 결과를 변경할 수 있는 입력 섭동의 양을 최소화하기 위해 최적화 기반 공격 방법이 제안되었으며, 주어진 모델에서 얻은 기울기(gradient)의 부호(sign)에서 섭동을 계산하는 빠른 기울기 부호 방법(FGSM, Fast Gradient Sing Method)이 개발되었다. 또한, Iterative FGSM(I-FGSM)은 FGSM을 반복 접근 방식으로 확장하였으며 FGSM보다 공격의 속임률을 더 높일 수 있다. 이러한 방법은 주어진 입력 이미지의 전체 영역에서 섭동을 찾는 반면, 다른 방법은 심층 이미지 분류기(deep image classifier)를 속이기 위해 단 하나의 픽셀의 섭동을 찾고자 하는 단일 픽셀 공격(one-pixel attack)의 가능성을 보여주었다.

[0044] 딥러닝 모델의 취약성(vulnerability)은 각 입력 이미지에 대한 섭동을 찾는 것 이상의 여러 고급 방법을 통해 추가로 평가될 수 있다. 특정 방법은 하나의 모델에서 발견된 섭동이 다른 모델에서도 작동할 수 있는지 여부를 조사하는 섭동의 전이성(transferability)을 조사하였으며, 다른 방법은 주어진 분류기를 속이기 위해 모든 이미지에 적용할 수 있는 이미지 독립적인 보편적 섭동(universal perturbation)을 발견하였다.

[0045] 행동인식 모델의 화이트 박스 적대적 공격(white-box adversarial attack)에 대한 몇 가지 연구가 수행되어 왔다. 제1 연구는 GAN(Generative Adversarial Network)을 사용하여 합성곱 3D(C3D) 모델(convolutional 3D model)에 대한 적대적 공격을 개발하였다. 제2 연구는 LSTM 기반 모델에 대한 적대적 섭동(adversarial perturbation)을 생성하기 위해 최적화 기반 방법을 제안하였다. 제3 연구는 깜박거리는 섭동(flickering perturbation)을 얻기 위해 주어진 비디오 클립에서 각 프레임의 전체 색상을 변경하는 방법을 개발하였다. 그러나, 이러한 방법들은 기존의 행동인식 모델에서만 검증되었으며, 주어진 비디오의 여러 프레임들에 섭동을 추가하기 때문에 인간 관찰자(human observer)에 의해 감지될 가능성이 존재할 수 있다.

[0046] 이하, 도 1 내지 11을 참조하여 본 발명에 대해 보다 자세히 설명한다.

[0047] 도 1은 본 발명에 따른 적대적 공격 장치의 물리적 구성을 설명하는 도면이다.

[0048] 도 1을 참조하면, 적대적 공격 장치(100)는 정상인식 구간 결정부(110), 취약 프레임 검출부(120), 공격 프레임 생성부(130), 공격 성능 분석부(140) 및 제어부(도 1에 미도시함)를 포함할 수 있다.

[0049] 정상인식 구간 결정부(110)는 대상 행동인식 모델을 통해 정상적으로 인식되는 오리지널 비디오 클립의 특정 구간을 결정할 수 있다. 여기에서, 오리지널 비디오 클립(original video clip)은 동영상 파일에 해당할 수 있으며, 복수의 이미지 프레임들로 구성될 수 있다. 특히, 오리지널 비디오 클립은 기 구축된 대상 행동인식 모델(target action recognition model)을 통해 정상적으로 인식되는 원본의 비디오 클립에 해당할 수 있다. 정상인식 구간 결정부(110)는 입력으로 주어진 오리지널 비디오 클립을 적대적 공격 장치(100)와 연결된 사용자 단말을 통해 수신할 수 있으며, 데이터베이스에 저장된 오리지널 비디오 클립을 읽어올 수 있다. 또한, 정상인식 구간 결정부(110)는 오리지널 비디오 클립에서 소정의 구간을 특정할 수 있으며, 해당 특정 구간은 이후 단계에서 적대적 공격을 위한 대상 비디오 클립으로 사용될 수 있다.

[0050] 일 실시예에서, 정상인식 구간 결정부(110)는 오리지널 비디오 클립의 제1 재생시점부터 제2 재생시점까지의 연속적인 프레임들로 특정 구간을 구성할 수 있다. 즉, 정상인식 구간 결정부(110)는 대상 행동인식 모델의 취약성 분석을 위해 오리지널 비디오 클립 상에서 시간적으로 연속하는 프레임들을 특정 구간으로 결정할 수 있다. 특정 구간의 시작점은 제1 재생시점에 해당할 수 있으며, 종료점은 제2 재생시점에 해당할 수 있다. 따라서, 제1 및 제2 재생시점 사이에 연속적으로 존재하는 프레임들이 특정 구간으로서 결정될 수 있다. 특정 구간 내의 프레임들은 기본적으로 대상 행동인식 모델에 의해 정상적으로 인식되는 프레임들에 해당할 수 있다.

[0051] 취약 프레임 검출부(120)는 특정 구간에서 원본 프레임과의 차이를 특정 기준 이하로 유지시키는 섭동(perturbation)을 통해 대상 행동인식 모델의 오인식을 발생시키는 적어도 하나의 취약 프레임을 검출할 수 있다. 즉, 취약 프레임 검출부(120)는 대상 행동인식 모델의 구조적 취약성을 분석하기 위하여 비디오 클립 상의 소정의 프레임에 작은 변화를 가하는 섭동을 주입시킬 수 있으며, 이를 기초로 대상 행동인식 모델의 성능을 분석하여 구조적 취약성을 인지할 수 있다. 결과적으로, 취약 프레임 검출부(120)에 의해 검출된 취약 프레임에 대해 섭동을 중점적으로 주입하는 경우 대상 행동인식 모델의 오인식을 유발시킬 수 있으며, 이를 통해 적대적 공격을 달성할 수 있다.

[0052] 일 실시예에서, 취약 프레임 검출부(120)는 연속적인 프레임들 중 어느 하나에 섭동으로서 균일 랜덤 노이즈(uniform random noise)를 주입할 수 있다. 여기에서, 균일 랜덤 노이즈 방식은 소정의 범위 내에서 노이즈를 랜덤하게 선택하여 섭동으로서 프레임에 주입하는 방식에 해당할 수 있다. 취약 프레임 검출부(120)는 연속적인 프레임들 중에서 선택된 프레임에 대해 균일 랜덤 노이즈 방식을 통해 섭동을 주입시킬 수 있다.

[0053] 일 실시예에서, 취약 프레임 검출부(120)는 오리지널 비디오 클립을 대상으로 대상 행동인식 모델의 구조적 취약성(structural vulnerability)을 분석하기 위하여 I-FGSM 알고리즘을 적용하여 연속적인 프레임들 중 어느 하나에 섭동을 주입할 수 있다. I-FGSM 알고리즘은 FGSM 알고리즘을 확장한 버전에 해당할 수 있으며, 반복적인 과정을 통해 연속적인 프레임들 중에서 특정 프레임을 선택하고 기 정의된 섭동을 주입시킬 수 있다.

[0054] 일 실시예에서, 취약 프레임 검출부(120)는 I-FGSM 알고리즘의 반복마다 다음의 수학식 1을 통해 연속적인 프레임들에서 i번째 프레임을 검출할 수 있다.

[0055] [수학식 1]

[0056] $$\tilde{X}^{n+1}(i) = Clip_{0,255}\left(X^n(i) + \frac{\epsilon}{N} sgn\left(\nabla_{X^n(i)} J(X^n, y)\right)\right)$$

[0057] $$X^{n+1}(i) = Clip_{-\epsilon,\epsilon}\left(\tilde{X}^{n+1}(i) - X^0(i)\right) + X^0(i)$$

[0059] 여기에서, ϵ는 추가될 섭동량(amount of perturbation)이고, sgn(·)은 부호 함수이고(sign function), $\nabla_{X^n(i)} J(X^n, y)$는 손실 함수(loss function)에 대한 대상 프레임의 그라디언트(gradient)이다.

[0060] 일 실시예에서, 취약 프레임 검출부(120)는 복수의 오리지널 비디오 클립들에 대한 보편적 섭동(universial perturbation)을 통해 적어도 하나의 취약 프레임을 검출할 수 있다. 취약 프레임 검출부(120)는 기본적으로 각 오리지널 비디오 클립마다 취약 프레임에 관한 적대적 섭동을 찾을 수 있다. 또한, 취약 프레임 검출부(120)는 복수의 오리지널 비디오 클립들에 대해 비디오 독립적인 섭동(video-agnostic perturbation)을 찾을 수 있다.

이를 위해, 취약 프레임 검출부(120)는 I-FGSM 알고리즘과 유사한 방법을 적용할 수 있다. 즉, 취약 프레임 검출부(120)는 각 비디오 클립에 대한 그라디언트 대신 복수의 비디오 클립들에 대한 평균 그라디언트(average gradient)를 부호 함수(sign function)에 적용할 수 있으며, 비디오 클립에 독립적인 보편적 섭동을 획득할 수 있다.

[0061] 일 실시예에서, 취약 프레임 검출부(120)는 섭동이 주입된 프레임을 대상 행동인식 모델에 입력하여 섭동에 관한 해당 프레임의 속임률(Fooling rate)을 산출할 수 있다. 취약 프레임 검출부(120)는 대상 행동인식 모델을 통해 섭동이 주입된 프레임에 대한 행동인식 동작을 수행할 수 있으며, 행동인식 결과를 기초로 섭동에 따른 속임률을 산출할 수 있다.

[0062] 일 실시예에서, 취약 프레임 검출부(120)는 연속적인 프레임들 각각에 대해 속임률을 산출하고 속임률을 기초로 적어도 하나의 취약 프레임을 결정할 수 있다. 즉, 속임률이 상대적으로 높은 프레임일수록 적대적 공격에 취약할 수 있으며, 속임률이 특정 값을 초과하는 경우 취약 프레임으로 결정될 수 있다. 취약 프레임 검출부(120)는 프레임 인덱스(frame index)를 기초로 반복적인 과정을 통해 프레임별 속임률을 산출할 수 있고, 속임률을 기설정된 값과 비교하여 적어도 하나의 취약 프레임을 결정할 수 있다. 취약 프레임은 섭동의 주입만으로 대상 행동인식 모델이 오인식하도록 하는 점에서 대상 행동인식 모델의 구조적 취약성에 큰 영향을 미칠 수 있다.

[0063] 공격 프레임 생성부(130)는 적어도 하나의 취약 프레임에만 섭동을 주입하는 화이트박스 시나리오에 따라 적어도 하나의 취약 프레임에 대응되는 공격 프레임을 생성할 수 있다. 즉, 공격 프레임 생성부(130)는 대상 행동인식 모델을 공격하기 위하여 비디오 클립 상에서 취약 프레임이 검출된 경우 해당 취약 프레임을 대상으로 섭동을 주입하는 적대적 공격(adversarial attack)을 통해 공격 프레임을 생성할 수 있다. 취약 프레임과 공격 프레임 간의 차이는 눈에 띄지 않는 정도인 점에서 공격 프레임을 포함하는 비디오 클립에 대해 대상 행동인식 모델은 잘못된 인식 결과를 생성할 수 있다.

[0064] 일 실시예에서, 공격 프레임 생성부(130)는 적어도 하나의 취약 프레임 중에서 속임률이 가장 높은 프레임에 대해 단일 프레임 공격을 통해 공격 프레임을 생성할 수 있다. 하나의 오리지널 비디오 클립 상에서 복수의 취약 프레임들이 존재할 수 있으며, 공격 프레임 생성부(130)는 복수의 취약 프레임들 중에서 가장 높은 속임률과 연관된 프레임만을 대상으로 적대적 공격을 통해 섭동이 주입된 공격 프레임을 생성할 수 있다. 대상 행동인식 모델의 구조적인 취약성을 이용하는 경우 단일 프레임 공격이 기존의 공격 방법들에 비해 높은 속임률과 높은 비가시성(invisibility)을 제공할 수 있음이 하기에서 설명하는 실험을 통해 입증될 수 있다.

[0065] 공격 성능 분석부(140)는 오리지널 비디오 클립의 특정 구간에 공격 프레임을 대체 삽입하여 대상 행동인식 모델에 대한 공격 성능을 평가할 수 있다. 취약 프레임 검출부(120)를 통해 검출된 취약 프레임에 대해 공격이 진행된 결과 취약 프레임이 공격 프레임으로 변경될 수 있으며, 공격 성능 분석부(140)는 공격된 비디오 클립을 대상으로 대상 행동인식 모델의 성능을 측정함으로써 공격 성능을 평가할 수 있다.

[0066] 일 실시예에서, 공격 성능 분석부(140)는 속임률(fooling rate)과 섭동의 눈에 띄지 않는 정도(inconspicuousness)를 기초로 공격 성능을 평가할 수 있다. 이에 대해서는 하기의 실험을 통해 보다 구체적으로 설명한다.

[0067] 제어부(도 1에 미도시함)는 적대적 공격 장치(100)의 전체적인 동작을 제어하고, 정상인식 구간 결정부(110), 취약 프레임 검출부(120), 공격 프레임 생성부(130) 및 공격 성능 분석부(140) 간의 제어 흐름 또는 데이터 흐름을 관리할 수 있다.

[0069] 도 2는 본 발명에 따른 행동인식을 위한 인지 최적화 적대적 공격 방법을 설명하는 순서도이다.

[0070] 도 2를 참조하면, 적대적 공격 장치(100)는 정상인식 구간 결정부(110)를 통해 대상 행동인식 모델을 통해 정상적으로 인식되는 오리지널 비디오 클립의 특정 구간을 결정할 수 있다(단계 S210).

[0071] 적대적 공격 장치(100)는 취약 프레임 검출부(120)를 통해 특정 구간에서 원본 프레임과의 차이를 특정 기준 이하로 유지시키는 섭동(perturbation)을 통해 대상 행동인식 모델의 오인식을 발생시키는 적어도 하나의 취약 프레임을 검출할 수 있다(단계 S220).

[0072] 적대적 공격 장치(100)는 공격 프레임 생성부(130)를 통해 적어도 하나의 취약 프레임에만 섭동을 주입하는 화이트박스 시나리오에 따라 적어도 하나의 취약 프레임에 대응되는 공격 프레임을 생성할 수 있다(단계 S230).

[0074] 이하, 도 3 내지 11을 참조하여, 본 발명에 따른 행동인식을 위한 딥러닝 모델에 대한 인지 최적화 적대적 공격

장치 및 방법에 대해 보다 자세히 설명한다.

[0075] 본 발명에 따른 적대적 공격 방법(이하, 적대적 공격 방법)은 행동인식 모델의 구조적 취약성을 분석하여 공격할 수 있다. 행동인식 모델의 구조적 취약성을 분석하기 위하여 비디오 시퀀스의 단일 프레임은 I-FGSM 알고리즘과 균일 랜덤 노이즈에 의해 섭동될 수 있고, 각 프레임에 대한 인식 성능이 평가될 수 있다. 이와 함께, 해당 취약성을 유발하는 요인이 분석될 수 있다.

[0076] 적대적 공격 방법은 I-FGSM 알고리즘을 사용하여 프레임을 섭동할 수 있으며, I-FGSM 알고리즘은 널리 사용되는 강력한 적대적 공격 방법 중 하나에 해당할 수 있다.

[0077] 적대적 공격 방법은 I-FGSM 알고리즘을 통해 비디오 클립에서 i번째 프레임의 섭동을 반복적으로 찾는 동작을 수행할 수 있다. X^0={X^0(1), X^0(2), ..., X^0(T)}는 대상 행동인식 모델 M(·)에 의해 y로 정상 분류된 오리지널 비디오 클립(T개의 프레임을 가짐)에 해당한다. 즉, M(X^0) = y이다. 공격은 X^0로부터 비디오 클립 X의 공격된 버전(attacked version)을 찾는 것을 목표로 하며, 여기에서 i번째 프레임 X^0(i)만 눈에 띄지 않는 섭동을 포함하는 공격된 X(i) 프레임으로 변경될 수 있다. 구체적으로, X(i)를 찾기 위해 I-FGSM 갱신 규칙(update rule)이 적용될 수 있다. 즉, 이전 반복의 프레임 X^n(i)로부터 n+1 번째 반복의 적대적 프레임 X^{n+1}(i)을 반복적으로 찾을 수 있으며, 다음의 수학식 1 및 2와 같이 표현될 수 있다.

[0078] [수학식 1]

[0079] $$\tilde{X}^{n+1}(i) = Clip_{0,255}\left(X^n(i) + \frac{\epsilon}{N} sgn\left(\nabla_{X^n(i)} J(X^n, y)\right)\right)$$

[0080] [수학식 2]

[0081] $$X^{n+1}(i) = Clip_{-\epsilon,\epsilon}\left(\tilde{X}^{n+1}(i) - X^0(i)\right) + X^0(i)$$

[0083] 여기에서, ϵ는 추가될 섭동량(amount of perturbation)을 조절하고, sgn(·)은 부호 함수(sign function)이고, $\nabla_{X^n(i)} J(X^n, y)$는 손실 함수(loss function) $J(X^n, y)$에 대한 대상 프레임의 그라디언트이다. 또한, $Clip_{a,b}(X)$는 다음의 수학식 3과 같이 표현될 수 있다.

[0084] [수학식 3]

[0085] $$Clip_{a,b}(X) = \min\left(\max(X, a), b\right)$$

[0087] 또한, N번의 반복 이후 최종적인 적대적 비디오 클립은 X = X^N = {X^0(1), X^0(2), ..., X^N(i), X^0(i+1), ..., X^0(T)}에 의해 획득될 수 있다. 결과적으로, X가 행동인식 모델에 입력되는 경우 행동인식 모델은 잘못된 예측을 출력할 할 수 있다. 즉, M(X) ≠ y이다.

[0088] 또한, 적대적 공격 방법은 [-64, 64] 내의 균일한 랜덤 노이즈를 섭동으로 사용하여 특정 프레임에 주입할 수 있다. 이러한 유형의 섭동은 구조적 취약성을 분석하기 위해 테스트될 수 있다. 또한, 랜덤 노이즈 섭동을 생성하는 것이 계산적으로 효율적이기 때문에 주어진 행동인식 모델에 대해 취약한 프레임 인덱스들(frame indices)을 식별하는데 사용될 수 있다.

[0089] 적대적 공격 방법에 관한 행동인식 모델의 구조적 취약성을 분석하기 위하여 소정의 실험이 수행될 수 있다. 해당 실험에서는 행동인식을 위해 널리 사용되는 대규모 벤치마크 데이터셋 중 하나인 Kinetics-400이 사용될 수 있다. 먼저, Kinetics-400의 테스트셋에서 각 클래스에 대해 10개의 비디오들이 무작위로 선택될 수 있다. 이에 따라, 적대적 공격 방법의 속임률(fooling rate)을 평가하기 위해 총 4000개의 비디오가 선택될 수 있다. 대상 행동인식 모델에는 I3D, SlowFast, ir-CSN 등 다양한 모델 구조를 갖는 3가지 최신 모델이 사용될 수 있다. 해당 모델들은 Kinetics-400 데이터셋에서 뛰어난 인식 성능을 나타낼 수 있다. 또한, MMAction2에서 사용할 수 있는 사전 학습된 모델이 사용될 수 있으며, MMAction2는 행동인식 모델에 대한 테스트 도구를 제공하는 오픈 소스 저장소에 해당할 수 있다. 이때, SlowFast 모델의 다양한 변형들 중에서 MMAction2의 8×8 SlowFast이 사

용될 수 있다.

[0090] 또한, 해당 실험에서는 다양한 하이퍼파라미터를 적용하여 I-FGSM 방법을 실행시킬 수 있다. 단일 프레임을 공격하기 위해 반복 횟수는 N ∈ {30, 50, 100}으로 설정될 수 있고, 섭동량(amount of perturbation)은 ϵ ∈ {2, 4, 8, 16}로 설정될 수 있다. 또한, 해당 실험에서는 경험적으로 대상 모델을 공격하기에 충분한 것으로 확인된 N = 30인 경우만이 관찰될 수 있다.

[0091] 도 3 및 4를 참조하면, 세가지 모델들에 대해 두 가지 유형의 섭동에 관한 속임률이 도시되어 있다. 특히, I3D 및 SlowFast의 경우 다른 것들보다 훨씬 더 높은 속임률을 나타내는 취약한 프레임 인덱스(또는 간단히 취약 프레임)의 존재가 관찰될 수 있다. 또한, 이러한 취약 프레임이 주기적으로 관찰될 수 있다. 구체적으로, I3D와 SlowFast는 각각 i ∈ {3, 7, 11, 15, 19, 23, 27, 31} 및 i ∈ {1, 5, 9, 13, 17, 21, 25, 29}에서 취약 프레임들이 나타날 수 있다. 다만, ir-CSN은 해당 경향이 관찰되지 않을 수 있다. ir-CSN은 다른 두 모델에 비해 프레임 간 편차가 비교적 적을 수 있으며, 전반적으로 상대적으로 높은 취약성을 나타낼 수 있다. 또한, 이러한 관찰 내용은 ϵ의 다른 값들에도 일관되게 적용될 수 있다.

[0092] 취약 프레임들 중 가장 취약한 프레임은 I3D, SlowFast, ir-CSN에 대해 각각 31번째, 29번째, 1번째 프레임일 수 있다. 균일한 랜덤 노이즈를 추가하더라도 가장 취약한 프레임을 식별할 수 있다. 도 4의 경우 모든 비디오 클립을 사용한 결과로서 획득되었으나, 무작위로 선택한 100개의 비디오 클립만으로도 가장 취약한 프레임을 발견하기에 충분할 수 있다.

[0093] 도 5를 참조하면, 행동인식 모델의 구조적 특성을 분석하여 취약 프레임에 관한 관찰의 원인을 도출할 수 있다.

[0094] 도 5의 그림 (a)에서, I3D 모델은 커널 크기(kernel size)가 5×7×7이고 시간 스트라이드(temporal stride)가 2인 컨볼루션 레이어(convolutional layer)를 통해 처음에 주어진 비디오 클립으로부터 피처(feature)들을 추출할 수 있다. 그런 다음, 해당 피처들은 커널 크기가 1×3×3이고 시간 스트라이드가 2인 맥스 풀링 레이어(max-pooling layer)에 의해 처리될 수 있다. 이 과정을 통해, 32 프레임의 비디오 클립은 8 프레임의 피처들로 축약(contract)될 수 있다. 즉, 첫번째 컨볼루션 레이어와 맥스 풀링 레이어에 의한 시간 스트라이드는 4에 해당한다고 할 수 있다. 그리고, 두 레이어를 통해 비대칭 정보의 추출이 존재할 수 있다. 즉, 컨볼루션 레이어의 가중치가 시간적 차원(temporal dimension) 전체에 걸쳐 서로 다른 값을 갖는 경우, 해당 레이어의 출력은 특정 프레임의 정보에 더 의존적일 수 있고, 커널로 진입하는 5개의 프레임들 중 다른 프레임의 정보에 덜 의존적일 수 있으며, 맥스 풀링 레이어를 통해 더욱 강조될 수 있다. 해당 실험에서 사용된 사전 학습된 I3D 모델의 경우, 첫 번째 컨볼루션 레이어의 가중치들의 평균 크기는 0.01, 0.01, 0.02, 0.03, 0.10으로 측정될 수 있다. 즉, 해당 레이어는 5개의 입력 프레임들 중 5번째 프레임의 정보에 대부분 의존적일 수 있다(도 5의 그림 (a)에서 두꺼운 선으로 표시됨). 해당 두 가지 메커니즘(즉, 4의 효과적인 스트라이드와 비대칭 정보의 추출)에 의해, i ∈ {3, 7, 11, 15, 19, 23, 27, 31}에 삽입된 섭동은 다른 프레임에서의 섭동과 달리 행동인식 모델을 쉽게 공격할 수 있다(도 5의 그림 (a)에서 빨간색 상자로 표시된 프레임).

[0095] 도 5의 그림 (b)에서, SlowFast는 빠른 경로(Fast Pathway)와 느린 경로(Slow Pathway)를 포함하는 2-스트림 모델(two-stream model)에 해당할 수 있다. 빠른 경로는 32프레임 모두를 사용하고 느린 경로는 매 4번째 프레임만 사용하기 때문에, 두 경로는 모두 8프레임만 동시에 사용할 수 있다(도 5의 그림 (b)에 빨간색 상자로 표시). 매우 취약한 프레임들(i ∈ {1, 5, 9, 13, 17, 21, 25, 29})은 두 경로에서 사용하는 프레임과 정확히 일치할 수 있다. 다른 프레임들은 빠른 경로를 통해서만 처리되므로 해당 프레임들의 섭동은 상대적으로 성공적이지 않을 수 있다. 빠른 경로는 I3D의 구조를 갖지만 원래 I3D와 달리 시간 스트라이드는 1이므로 위에서 관찰되는 주기적인 패턴이 여기에서는 나타나지 않을 수 있다.

[0096] 또한, ir-CSN 모델은 ResNet-152를 기반으로 하며, 이는 ResNet-50을 사용하는 다른 두 모델보다 더 깊은 모델에 해당할 수 있다. 앞서 언급했듯이, ir-CSN 모델은 모든 프레임에서 상대적으로 취약할 수 있다. 예를 들어, ϵ가 16일 때 가장 낮은 속임률이 75.4%일 수 있다. 첫 번째 컨볼루션 레이어와 첫 번째 풀링 레이어의 스트라이드는 1이므로 모든 입력 프레임은 해당 레이어들에서 균등하게 처리될 수 있다. 따라서, 취약성은 I3D 및 SlowFast와 달리 모든 프레임에서 다소 유사하게 나타날 수 있다. 또한, 두 개의 에지 프레임들(edge frames) 주변에서 속임률이 증가하는 것으로 관찰될 수 있으며, 이는 제로 패딩(zero paddings)이 에지 프레임의 섭동을 강조(highlight)하기 때문일 수 있다.

[0097] 또한, 해당 실험에서는 I-FGSM을 사용하여 프레임 간 섭동의 전이성(transferability), 즉 특정 프레임에 대해

생성된 섭동이 다른 프레임 위치에도 직접 사용되어 행동인식 모델을 공격할 수 있는지 여부가 조사될 수 있다. 도 6에서, 특히 높은 전이성을 보이는 소스 프레임 위치(source frame location)와 타겟 프레임 위치(target frame location)의 쌍이 존재할 수 있으며, 전이성 패턴(transferability)은 행동인식 모델에 따라 달라질 수 있다.

[0098] I3D 및 SlowFast의 경우 상대적으로 더 취약한 프레임들 간에 보다 높은 전이성이 도출될 수 있다(도 3 참조). 전이된 섭동에 의한 속임률은 균일 랜덤 노이즈 공격에 의한 속임률보다 높게 나타날 수 있다. 이는 취약한 프레임에 대한 섭동이 행동인식 모델을 잘못 작동시키는 공통 피처들(common features)을 가지고 있음을 나타낼 수 있다.

[0099] 더욱이, I3D와 ir-CSN의 경우, 인접한 프레임 간에는 상대적으로 높은 수준의 전이성이 나타날 수 있다. 그러나, SlowFast에서는 느린 경로가 인접한 모든 프레임을 사용하는 것이 아니라 네 번째 프레임만 사용하기 때문에 그렇지 않을 수 있다.

[0100] 적대적 공격 방법은 I-FGSM 알고리즘을 가장 취약한 프레임(즉, I3D, SlowFast 및 ir-CSN에 대해 31번째, 29번째 및 첫 번째 프레임)에만 적용하는 화이트 박스 시나리오에서 단일 프레임 공격(one frame attack)을 수행할 수 있다. 적대적 공격 방법은 1)속임률(fooling rate)과 2) 섭동의 눈에 띄지 않는 정도(degree of inconspicuousness of perturbation)라는 두 가지 기준에 따라 공격의 성능이 평가될 수 있다. 행동인식 모델의 구조적인 취약성을 이용한 단일 프레임 공격은 기존 공격 방식에 비해 높은 속임률과 높은 비가시성(invisibility)으로 행동인식 모델을 속일 수 있다.

[0101] 도 7은 각 사례에 대한 단일 프레임 공격의 속임률을 나타낼 수 있다. 해당 공격은 모든 대상 모델에 대해 높은 속임률을 달성할 수 있다. 특히, ϵ가 8과 같거나 클 경우, 속임률은 90%를 초과할 수 있다. 또한, ϵ가 2보다 작을 때도 60%를 초과할 수 있다.

[0102] 비교를 위해 또 다른 공격 방법을 구현하여 심층 행동인식 모델들을 공격할 수 있다. ϵ = 2인 ir-CSN의 경우를 제외하고 해당 방법의 속임률은 단일 프레임 공격보다 낮을 수 있다. 최대 ϵ = 16인 단일 프레임 공격이 상당히 눈에 띄지 않는다는 점을 고려하면 해당 방법은 행동인식 모델의 취약성을 효과적으로 포착하지 못할 수 있으며, 이는 LSTM 기반 모델의 특정 메커니즘(시간 정보 전파)을 활용하도록 설계되었기 때문일 수 있다.

[0103] 단일 프레임 공격의 눈에 띄지 않는 정도(level of inconspicuousness)를 알아보기 위해 주관 테스트(subjective test)가 수행될 수 있다. 4(비디오 개수) × 4($\epsilon \in \{2, 4, 8, 16\}$) × 3(대상 모델 개수) = 48개의 섭동된 비디오들이 사용될 수 있다. 여기서는, ITU-R BT.500-13 추천에 따른 주관 테스트에 필요한 참가자 수를 충족하는 15명의 참가자가 참여할 수 있다. 주관 테스트는 이중 자극 손상 척도(DSIS, Double-Stimulus Impairment Scale) 방법을 기반으로 할 수 있다. 즉, 참가자는 DSIS에 따라 중회색 이미지(mid-gray image)가 둘 사이에 표시되는 오리지널 비디오와 섭동 버전을 각각 3초 동안 순차적으로 시청할 수 있다. 이때, 비디오 쌍의 순서는 무작위로 바뀔 수 있다. 그런 다음, 참가자는 비디오 쌍의 차이를 인지하는지 여부에 응답할 수 있다. 참가자의 '베이스라인' 감지 성능을 획득하기 위해 오리지널 비디오 쌍도 포함될 수 있다. 한 이미지의 노출 시간이 2초로 설정된 섭동 프레임과 원본 버전의 쌍에 대해 동일한 절차가 반복될 수 있다. 비교를 위해 비디오 클립을 공격하기 위한 모든 프레임에 깜박임 섭동(flickering perturbation)을 추가하여 높은 수준의 눈에 띄지 않는 것을 목표로 하는 깜박임 공격(flickering attack)이 구현될 수 있고, 결과 비디오들도 평가될 수 있다.

[0104] 도 8은 $\epsilon \in \{2, 4, 8, 16\}$를 설정할 때 I3D에 대한 섭동 프레임의 예를 나타낼 수 있다. 또한, 이러한 프레임(이미지로 볼 때)과 이러한 프레임이 포함된 비디오(비디오로 볼 때)의 탐지율(detection rate)이 제시될 수 있다. ϵ의 높은 값에 대한 프레임들의 섭동은 참가자가 쉽게 찾을 수 있다. 그 결과 동영상으로 볼 때보다 이미지로 볼 때 더 높은 탐지율을 획득할 수 있다.

[0105] 도 9는 주관 테스트(subjective)의 전체 결과(overall result)를 나타낼 수 있다. 단일 프레임 공격에 의해 섭동된 프레임은 이미지로 볼 때 특히 ϵ값이 큰 경우 비교적 쉽게 감지될 수 있다. ϵ가 커질수록 영상의 탐지율이 높아지는 것은 자연스러운 현상일 수 있다. 그러나, 섭동된 프레임이 포함된 비디오는 거의 탐지되지 않아 동일

한 오리지널 비디오 쌍의 잘못된 탐지율('베이스라인')보다 훨씬 낮은 탐지율을 나타낼 수 있다. 이와 달리, 깜박임 공격은 쉽게 탐지될 수 있다. 이러한 결과를 통해 단일 프레임 공격의 눈에 띄지 않음이 도출될 수 있다.

[0106] 적대적 공격 방법은 범용 공격(universal attack)의 가능성을 포함할 수 있다. 즉, 적대적 공격 방법은 대상 행동인식 모델의 비디오 클립에 영향을 미칠 수 있는 비디오 독립적인 섭동(video-agnostic perturbation)을 찾을 수 있다. 또한, 실시간 행동인식 상황을 가정하여 시간 불변 범용 공격(time-invariant universal attack)까지 단일 프레임 공격을 확장할 수 있는 가능성에 대해서도 설명한다.

[0107] 비디오에 독립적인 보편적 섭동(video-agnostic universal perturbation)은 상기의 수학식 1 및 2에 의해 설명된 방법과 유사한 방법으로 획득될 수 있다. 그러나, 각 비디오 클립에 대한 그라디언트 대신 K개의 비디오들에 대한 평균 그라디언트(average gradient)가 부호 함수(sign function)에 사용될 수 있으며, 다음의 수학식 4와 같이 표현될 수 있다.

[0108] [수학식 4]

[0109] $$G^{n+1} = \frac{1}{K}\sum_{k=1}^{K}\sum_{i\in I}\nabla_{X_k(i)}J(X_k^n, y_k)$$

[0111] 여기에서, I는 보편적 섭동을 탐색하는데 사용되는 대상 프레임 인덱스들의 집합이고, $X_k(i)$는 k번째 비디오 클립의 i번째 프레임이며, y_k는 k번째 비디오 클립의 정답(ground-truth) 레이블이다.

[0112] 또한, 보편적 섭동을 찾기 위해 두 개의 다른 프레임 세트 I를 사용할 수 있다. 즉, 여러 개의 매우 취약한 프레임 세트와 가장 취약한 프레임 세트이다. N=100 및 $\epsilon \in \{32, 48, 64\}$로 설정될 수 있다. 비디오 특정 섭동(video specific perturbation)의 경우에 비해 보편적 섭동을 찾기 위해서는 더 많은 반복 횟수와 더 큰 ϵ 값이 필요할 수 있다. 게다가, 모든 반복에서 모든 대상 비디오에서 그라디언트를 계산해야 하기 때문에 보편적인 섭동을 찾기 위해서는 높은 계산 복잡성(computation complexity)이 필요할 수 있다. 보편적인 섭동을 찾기 위해 더 많은 수의 비디오를 사용하는 것은 더 많은 시간이 걸리지만 더 다양한 비디오에서 섭동이 발견되기 때문에 더 높은 속임률을 기대할 수 있다. 이를 실험하기 위하여 $K \in \{100, 200, 500, 1000, 1500\}$에 대해 보편적 섭동을 찾기 위해 비디오 수(K)를 변경시킬 수 있다.

[0113] 보편적인 섭동을 테스트하기 위해 Kinetics-400 데이터셋에서 보편적 섭동을 생성하는데 사용된 비디오와 다른 1000개의 추가 비디오가 무작위로 선택될 수 있다.

[0114] 먼저, 모든 K개의 비디오 클립들에 대해 가장 취약한 프레임(즉, 상기 수학식 4의 I가 하나의 프레임 인덱스만 갖는 경우)에서 보편적 섭동을 찾을 수 있다. 도 10의 그림(a)는 K값에 대한 범용 단일 프레임 공격의 속임률을 나타낼 수 있다. 즉, 일반적으로 비디오 개수를 늘리는 것이 더 높은 속임률을 달성하는데 도움이 된다는 것을 나타낼 수 있다. 1500개의 동영상이 사용될 때 $\epsilon = 32$인 공격은 모든 대상 행동인식 모델에 대해 80% 이상의 속임률을 달성하여 범용 단일 프레임 공격이 가능함을 나타낼 수 있다.

[0115] 또한, 가장 취약한 프레임에서 발견되는 보편적 섭동의 전이성(Transferability)이 도출될 수 있다. 이를 위해, 모든 대상 프레임 인덱스에 대해 해당 프레임에 동일한 보편적 섭동을 추가하고 속임률을 측정할 수 있다. 도 10의 그림 (b)는 I3D 및 SlowFast에 대한 결과를 나타낼 수 있다. 또한, 취약한 프레임 인덱스는 전이된 보편적 섭동에 대해 매우 취약함을 나타낼 수 있다.

[0116] 단일 프레임 공격은 가장 취약한 프레임만 이용하더라도 강력한 보편적 섭동을 찾을 수 있지만, 여러 취약한 프레임을 사용하면 더 강력한 보편적 섭동을 찾을 수 있다. 이를 평가하기 위해 모든 취약한 프레임을 사용하여 보편적인 섭동을 찾을 수 있다(즉, 상기 수학식 4의 I에 여러 프레임 인덱스가 있는 경우). 도 10의 그림 (c)는 발견된 섭동이 가장 취약한 프레임에 주입되었을 때의 결과를 나타낼 수 있다. 가장 취약한 프레임만을 사용하여 찾아낸 보편적인 공격의 결과(도 10의 그림 (a))와 비교할 때, 해당 결과는 여러 취약한 프레임에서 발견된 보편적 섭동이 더 강력함을 나타낼 수 있다. 예를 들어, 500개 또는 1000개의 비디오가 보편적 섭동을 찾는데 사용되면 I3D와 SlowFast 모두에 대한 속임률은 도 10의 그림 (c)의 ϵ 값에서 90%를 초과할 수 있다. 또한, 적은 수의 비디오를 사용하는 경우에도 충분히 강한 보편적 섭동을 찾을 수 있다. 예를 들어, 200개의 비디오만

사용되는 경우에도 $\epsilon \in \{48, 64\}$인 모든 경우에 대해 속임률은 90%를 초과할 수 있다.

[0117] 또한, 비디오 데이터가 지속적으로 생성되고 슬라이딩 윈도우를 사용하여 생성된 비디오 시퀀스로부터 인식을 위해 비디오 클립이 반복적으로 선택되는 실시간 인식 시나리오가 고려될 수 있다. 해당 시나리오에서는 두 가지 문제가 발생할 수 있다. 첫째, 입력된 비디오와 관련된 섭동을 발생시키기에는 시간이 충분하지 않을 수 있다. 따라서, 실시간 동작을 보장하기 위해 우선 순위를 계산한 보편적 섭동이 사용될 필요가 있다. 둘째, 인식을 위해 선택한 비디오 클립과 공격자가 관찰하는 비디오 클립 사이에는 알 수 없는 시간적 오프셋이 존재할 수 있다. 이러한 시나리오를 처리할 수 있는 공격 방법이 시간 불변 공격(time-invariant attack)에 해당할 수 있다. 여기에서는, 시간 불변 공격으로서 범용 단일 프레임 공격이 실현 가능한지를 살펴본다.

[0118] 도 5에서 도시된 바와 같이, I3D 및 SlowFast에 대해 취약한 프레임이 주기적으로 나타날 수 있다. 이를 활용하여 다음과 같이 시간 불변 범용 공격(time-invariant universal attack)을 설계할 수 있다. P는 취약한 프레임의 기간(period of vulnerable frame)을 나타내고 L은 비디오 클립의 길이(length of video clip)를 나타내며, I3D와 SlowFast의 경우 L은 각각 4와 32이다. 그런 다음, 각 P 프레임의 프레임 인덱스가 P로 나누었을 때 구별되는 나머지(distinct remainder)에 대응되는 방식으로 여러 취약한 프레임을 공격하여 생성되는 사전 계산된(pre-computed) 범용 단일 프레임 섭동을 매 L 프레임마다 P 프레임에 추가할 수 있다. 예를 들어, 나머지가 각각 1, 2, 3, 0인 1, 10, 19, 24번째 프레임은 섭동되며, 이는 다음 32프레임에 대해 반복될 수 있다. 이렇게 하면 비디오 시퀀스의 어떤 프레임을 인식하기 위한 비디오 클립의 시작 프레임으로 선택했는지에 관계없이 P 섭동된 프레임 중 하나가 항상 취약한 프레임 인덱스와 일치할 수 있다. 해당 프로세스의 특별한 예로서, 매 L 프레임에서 첫 번째 P 프레임들을 섭동할 수 있다. 예를 들어, 첫 번째부터 네 번째, 33번째부터 36번째 프레임들에 해당할 수 있다.

[0119] 도 11은 보편적 섭동을 찾는데 사용된 비디오의 개수와 관련된 공격 결과를 나타낼 수 있다. 1000개의 비디오를 사용할 경우 모든 경우에 70% 이상의 속임률이 초과할 수 있다. 따라서, 시간 불변 범용 공격에 대한 도전적인 실시간 시나리오에서도 행동인식 모델이 매우 취약하다는 것을 확인할 수 있다.

[0121] 본 발명에 따른 적대적 공격 방법은 심층 행동인식 모델의 구조적 취약성에 대한 심층 분석 결과가 적용될 수 있다. 즉, 심층 분석 결과는 주어진 비디오 클립에서 단일 프레임을 섭동시킨 결과를 바탕으로, 해당 취약성은 컨볼루션 레이어와 맥스 풀링 레이어의 스트라이드들, 입력된 프레임들의 불균일 사용(uneven use) 등의 구조적 특성에 기인하는 것임을 나타낼 수 있다. 또한, 심층 분석 결과는 주관 테스트를 통해 눈에 잘 띄지 않는 것으로 밝혀진 강력한 단일 프레임 공격의 가능성을 입증할 수 있다. 마지막으로, 심층 분석 결과는 다양한 공격 시나리오에서 높은 속임률을 보이는 보편적 섭동 가능성을 도출할 수 있다.

[0123] 도 12는 본 발명에 따른 적대적 공격 장치의 시스템 구성을 설명하는 도면이다.

[0124] 도 12를 참조하면, 적대적 공격 장치(100)는 프로세서(1210), 메모리(1230), 사용자 입출력부(1250) 및 네트워크 입출력부(1270)를 포함할 수 있다.

[0125] 프로세서(1210)는 본 발명의 실시예에 따른 행동인식을 위한 딥러닝 모델에 대한 인지 최적화 적대적 공격 프로시저를 실행할 수 있고, 이러한 과정에서 읽혀지거나 작성되는 메모리(1230)를 관리할 수 있으며, 메모리(1230)에 있는 휘발성 메모리와 비휘발성 메모리 간의 동기화 시간을 스케줄 할 수 있다. 프로세서(1210)는 적대적 공격 장치(100)의 동작 전반을 제어할 수 있고, 메모리(1230), 사용자 입출력부(1250) 및 네트워크 입출력부(1270)와 전기적으로 연결되어 이들 간의 데이터 흐름을 제어할 수 있다. 프로세서(1210)는 적대적 공격 장치(100)의 CPU(Central Processing Unit)로 구현될 수 있다.

[0126] 메모리(1230)는 SSD(Solid State Disk) 또는 HDD(Hard Disk Drive)와 같은 비휘발성 메모리로 구현되어 적대적 공격 장치(100)에 필요한 데이터 전반을 저장하는데 사용되는 보조기억장치를 포함할 수 있고, RAM(Random Access Memory)과 같은 휘발성 메모리로 구현된 주기억장치를 포함할 수 있다. 또한, 메모리(1230)는 전기적으로 연결된 프로세서(1210)에 의해 실행됨으로써 본 발명에 따른 적대적 공격 방법을 실행하는 명령들의 집합을 저장할 수 있다.

[0127] 사용자 입출력부(1250)은 사용자 입력을 수신하기 위한 환경 및 사용자에게 특정 정보를 출력하기 위한 환경을 포함하고, 예를 들어, 터치 패드, 터치 스크린, 화상 키보드 또는 포인팅 장치와 같은 어댑터를 포함하는 입력장치 및 모니터 또는 터치 스크린과 같은 어댑터를 포함하는 출력장치를 포함할 수 있다. 일 실시예에서, 사용자 입출력부(1250)은 원격 접속을 통해 접속되는 컴퓨팅 장치에 해당할 수 있고, 그러한 경우, 적대적 공격 장

치(100)는 독립적인 서버로서 수행될 수 있다.

[0128] 네트워크 입출력부(1270)은 네트워크를 통해 사용자 단말(1310)과 연결되기 위한 통신 환경을 제공하고, 예를 들어, LAN(Local Area Network), MAN(Metropolitan Area Network), WAN(Wide Area Network) 및 VAN(Value Added Network) 등의 통신을 위한 어댑터를 포함할 수 있다. 또한, 네트워크 입출력부(1270)는 데이터의 무선 전송을 위해 WiFi, 블루투스 등의 근거리 통신 기능이나 4G 이상의 무선 통신 기능을 제공하도록 구현될 수 있다.

[0130] 도 13은 본 발명에 따른 적대적 공격 시스템을 설명하는 도면이다.

[0131] 도 13을 참조하면, 적대적 공격 시스템(1300)은 사용자 단말(1310), 적대적 공격 장치(100) 및 데이터베이스(1330)를 포함할 수 있다.

[0132] 사용자 단말(1310)은 사용자에 의해 운용되는 단말 장치에 해당할 수 있다. 본 발명의 실시예에서 사용자는 하나 이상의 사용자로 이해될 수 있으며, 복수의 사용자들은 하나 이상의 사용자 그룹으로 구분될 수 있다. 또한, 사용자 단말(1310)은 적대적 공격 시스템(1300)을 구성하는 하나의 장치로서 적대적 공격 장치(100)와 연동하여 동작하는 컴퓨팅 장치에 해당할 수 있다. 예를 들어, 사용자 단말(1310)은 적대적 공격 장치(100)와 연결되어 동작 가능한 스마트폰, 노트북 또는 컴퓨터로 구현될 수 있으며, 반드시 이에 한정되지 않고, 태블릿 PC 등 포함하여 다양한 디바이스로도 구현될 수 있다. 또한, 사용자 단말(1310)은 적대적 공격 장치(100)와 연동하기 위한 전용 프로그램 또는 어플리케이션(또는 앱, app)을 설치하여 실행할 수 있다.

[0133] 적대적 공격 장치(100)는 본 발명에 적대적 공격 방법을 수행하는 컴퓨터 또는 프로그램에 해당하는 서버로 구현될 수 있다. 또한, 적대적 공격 장치(100)는 사용자 단말(1310)과 유선 네트워크 또는 블루투스, WiFi, LTE 등과 같은 무선 네트워크로 연결될 수 있고, 네트워크를 통해 사용자 단말(1310)과 데이터를 송·수신할 수 있다.

[0134] 또한, 적대적 공격 장치(100)는 관련 동작을 수행하기 위하여 독립된 외부 시스템(도 1에 미도시함)과 연결되어 동작하도록 구현될 수 있다. 예를 들어, 적대적 공격 장치(100)는 포털 시스템, SNS 시스템, 클라우드 시스템 등과 연동하여 다양한 서비스를 제공하도록 구현될 수 있다.

[0135] 데이터베이스(1330)는 적대적 공격 장치(100)의 동작 과정에서 필요한 다양한 정보들을 저장하는 저장장치에 해당할 수 있다. 예를 들어, 데이터베이스(1330)는 비디오에 관한 정보를 저장할 수 있고, 학습 데이터와 모델에 관한 정보를 저장할 수 있으며, 반드시 이에 한정되지 않고, 적대적 공격 장치(100)가 본 발명에 따른 적대적 공격 방법을 수행하는 과정에서 다양한 형태로 수집 또는 가공된 정보들을 저장할 수 있다.

[0136] 또한, 도 13에서, 데이터베이스(1330)는 적대적 공격 장치(100)와 독립적인 장치로서 도시되어 있으나, 반드시 이에 한정되지 않고, 논리적인 저장장치로서 적대적 공격 장치(100)에 포함되어 구현될 수 있음은 물론이다.

[0138] 상기에서는 본 발명의 바람직한 실시예를 참조하여 설명하였지만, 해당 기술 분야의 숙련된 당업자는 하기의 특허 청구의 범위에 기재된 본 발명의 사상 및 영역으로부터 벗어나지 않는 범위 내에서 본 발명을 다양하게 수정 및 변경시킬 수 있음을 이해할 수 있을 것이다.

부호의 설명

[0140] 100: 적대적 공격 장치

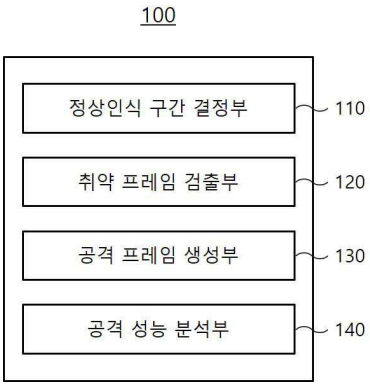
110: 정상인식 구간 결정부 120: 취약 프레임 검출부

130: 공격 프레임 생성부 140: 공격 성능 분석부

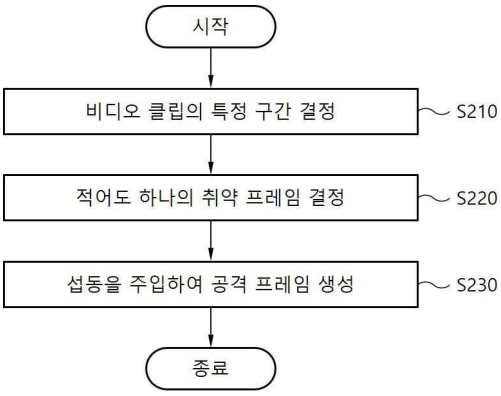
1300: 적대적 공격 시스템

도면

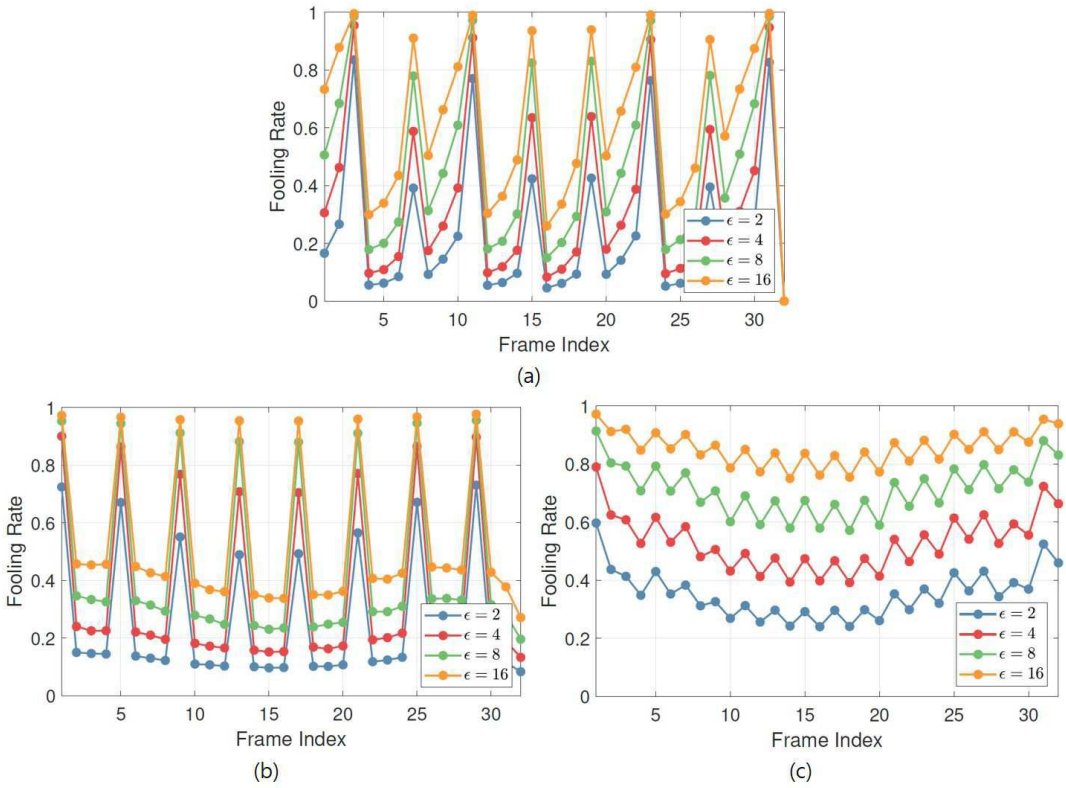
도면1


100
정상인식 구간 결정부
110
취약 프레임 검출부
120
공격 프레임 생성부
130
공격 성능 분석부
140

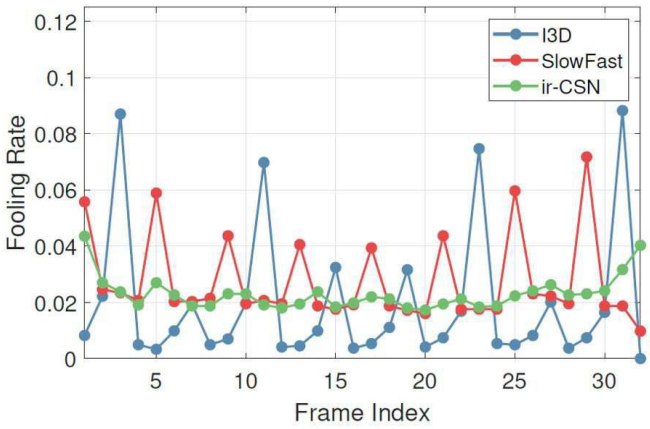
도면2


시작
비디오 클립의 특정 구간 결정
S210
적어도 하나의 취약 프레임 결정
S220
섭동을 주입하여 공격 프레임 생성
S230
종료

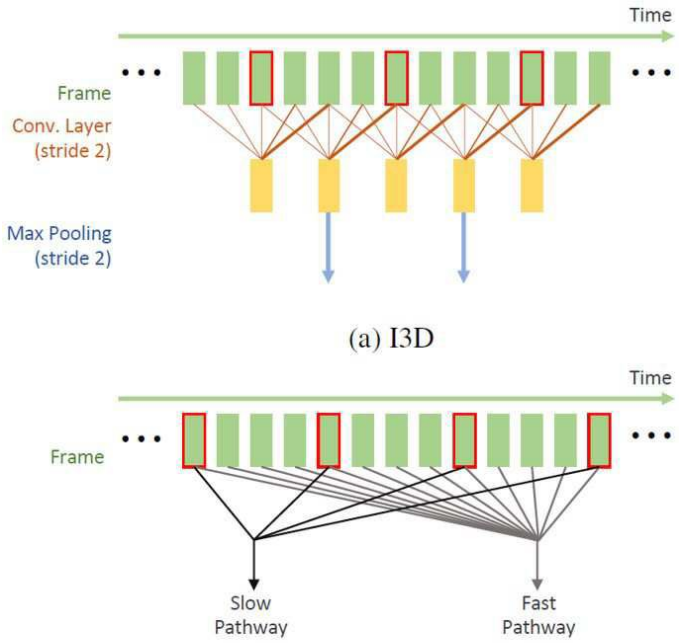
도면3


Fooling Rate
Frame Index
$\epsilon = 2$
$\epsilon = 4$
$\epsilon = 8$
$\epsilon = 16$
(a)
(b)
(c)

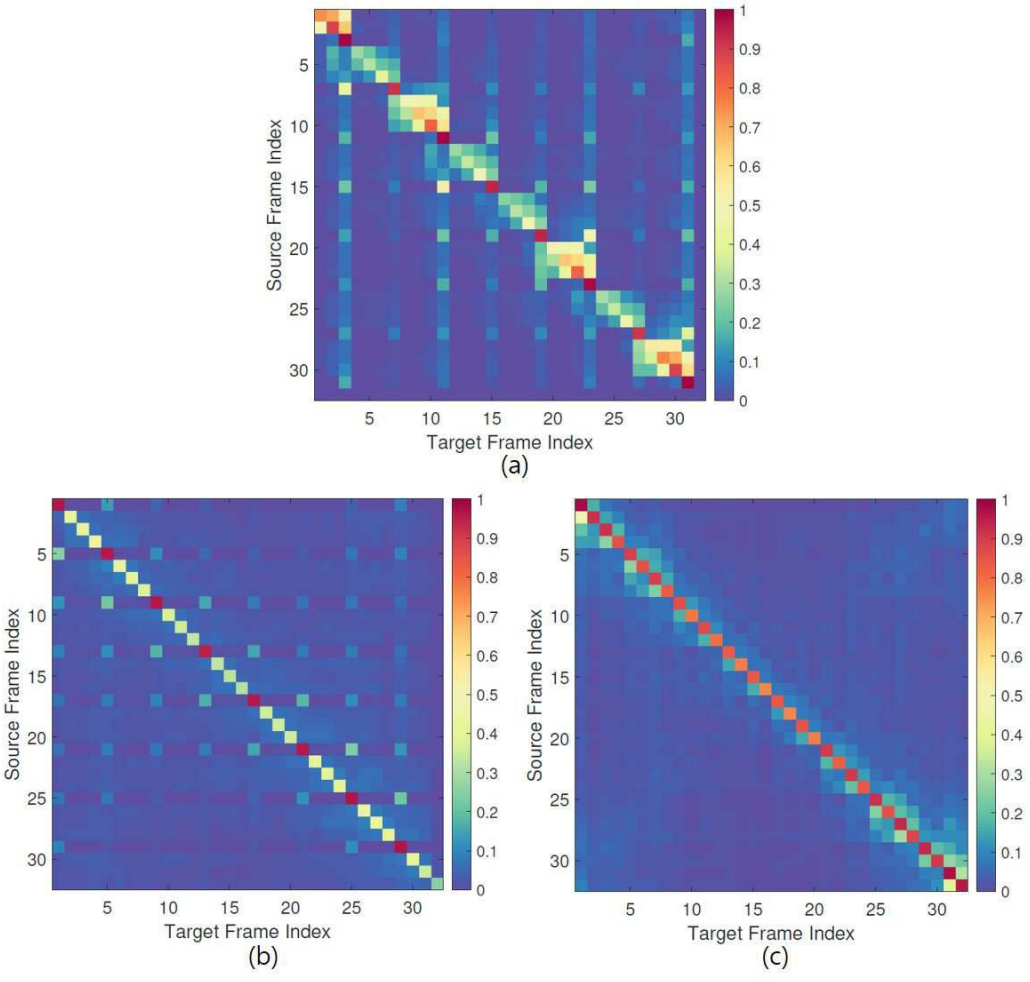
도면4


Fooling Rate
Frame Index
I3D
SlowFast
ir-CSN

도면5


Time
Frame
Conv. Layer
(stride 2)
Max Pooling
(stride 2)
(a) I3D
Time
Frame
Slow
Pathway
Fast
Pathway
(b) SlowFast

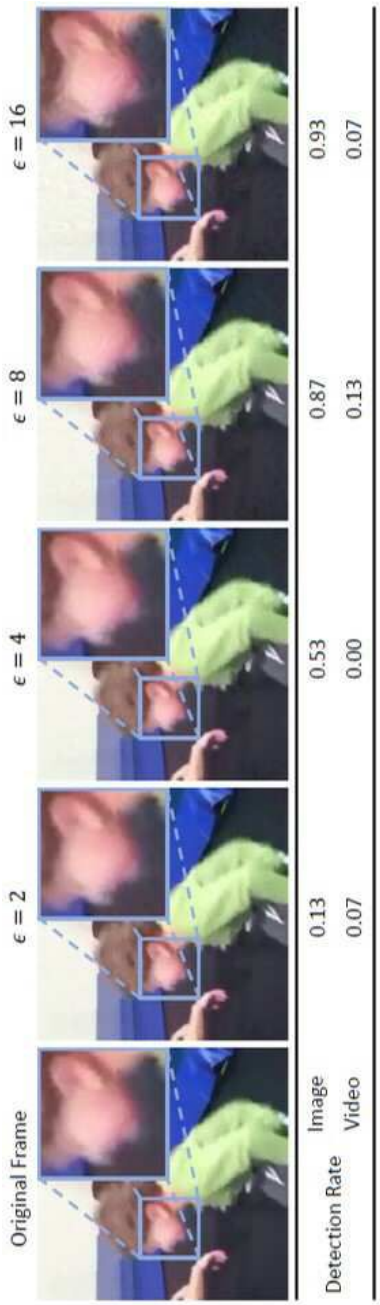
도면6


Source Frame Index
Target Frame Index
(a)
(b)
(c)

도면7

	One frame attack				Wei
	ϵ=2	ϵ=4	ϵ=8	ϵ=16	
I3D	0.83	0.95	0.99	1.00	0.81
SlowFast	0.73	0.90	0.96	0.98	0.72
ir-CSN	0.60	0.79	0.91	0.97	0.68

도면8



도면9

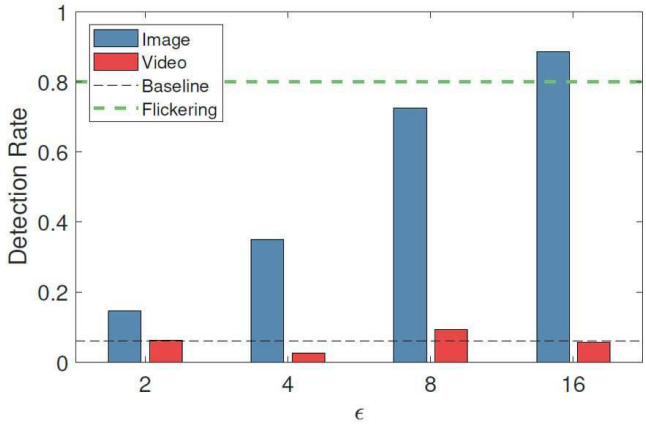
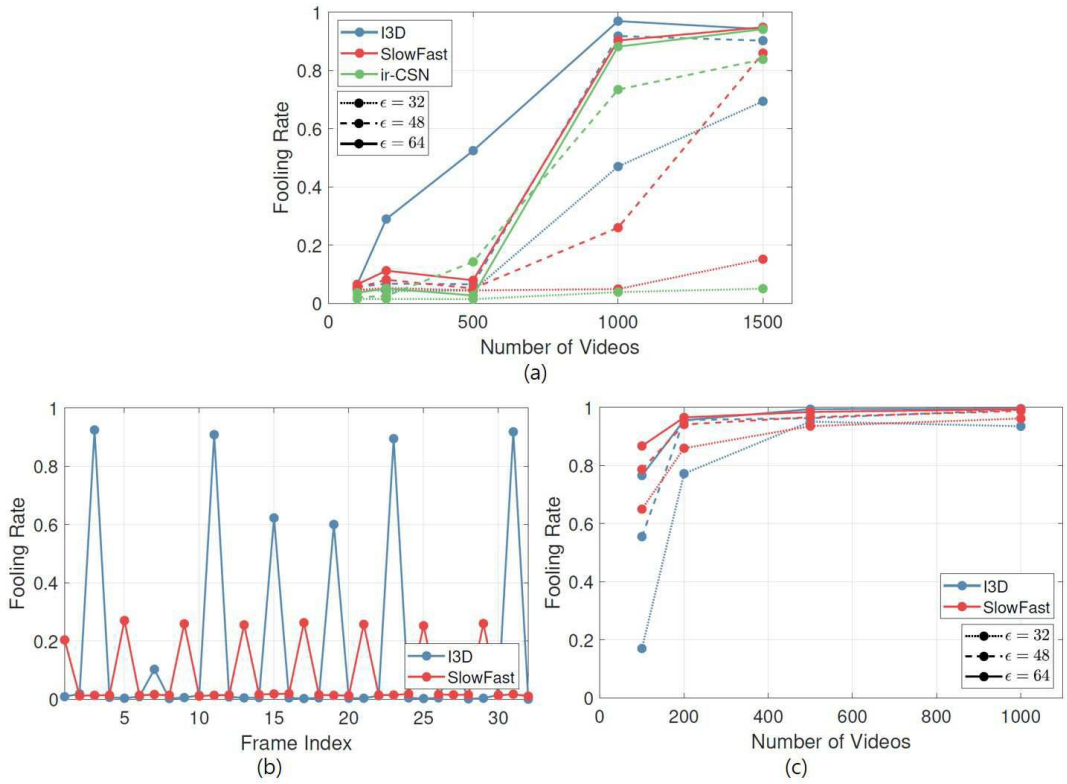
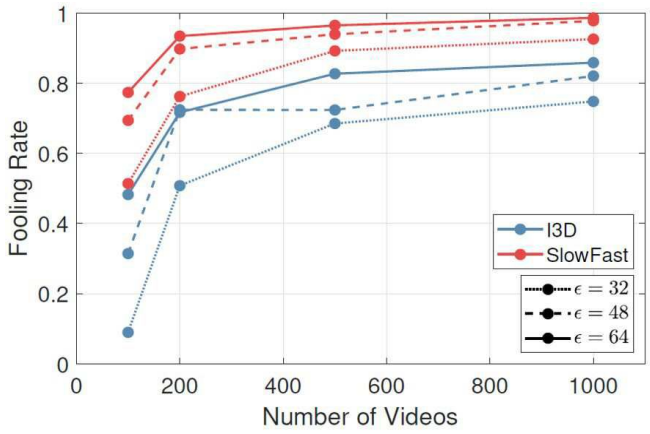


Image
Video
Baseline
Flickering
Detection Rate
1
0.8
0.6
0.4
0.2
0
2
4
8
16
ϵ

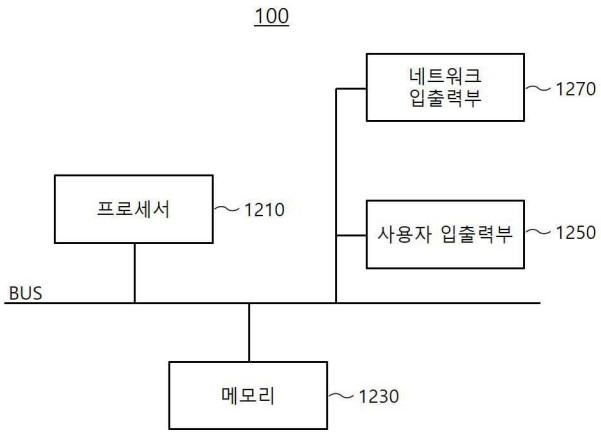
도면10


I3D
SlowFast
ir-CSN
ϵ = 32
ϵ = 48
ϵ = 64
Fooling Rate
Number of Videos
(a)
Fooling Rate
Frame Index
(b)
Fooling Rate
Number of Videos
(c)

도면11



도면12



도면13