적대적 공격과 뉴럴 렌더링 연구 동향 조사

이예진 [1], 심보석 [1], 허종욱 [1*]
[1] 한림대학교 소프트웨어학부
leeye0616@naver.com, bycicle55@naver.com, juhou@hallym.ac.kr

Survey Adversarial Attacks and Neural Rendering

Ye Jin Lee[1], Bo Seok Shim[1], Jong-Uk Hou[1*]
[1]Division of Software, Hallym University

요 약

다양한 분야에서 심층 신경망 기반 모델이 사용되면서 뛰어난 성능을 보이고 있다. 그러나 기계학습 모델의 오작동을 유도하는 적대적 공격(adversarial attack)에 의해 심층 신경망 모델의 취약성이 드러났다. 보안 분야에서는 이러한 취약성을 보완하기 위해 의도적으로 모델을 공격함으로써 모델의 강건함을 검증한다. 현재 2D 이미지에 대한 적대적 공격은 활발한 연구가 이루어지고 있지만, 3D 데이터에 대한 적대적 공격 연구는 그렇지 않은 실정이다. 본 논문에서는 뉴럴 렌더링(neural rendering)과 적대적 공격, 그리고 3D 표현에 적대적 공격을 적용한 연구를 조사해 이를 통해 추후 뉴럴 렌더링에서 일어나는 적대적 공격 연구에 도움이 될 것을 기대한다.

1. 서론

최근 컴퓨팅 기술의 발전과, 다양한 데이터를 보다 쉽게 수집할 수 있는 환경이 갖추어 짐에 따라 기계학습 기술은 최근 몇 년 동안 큰 발전 이루어지고 있으며 이미지 분류, 음성 인식, 자연어 처리 등의 분야에서 효과적인 작업을 수행해왔다. 2015 년에 ResNet [1]이 발표되면서, 이미지 인식 분야에서 기계학습이 사람의 성능을 초월하는 결과를 보여주었다. 인공지능 기술이 발전함에 따라 딥러닝은 자율 주행 기술과 같이 인지, 분류, 객체 탐지, 3D 표현, 위치 파악 등을 모두 다루는 다양한 문제들을 통합한 하나의 프레임워크로써 사용되고 있다

오늘날 메타버스, 가상/증강현실, 영화 등 각종 3D 분야의 연구가 활발히 이루어지고 있다. 3D 물체 생성의 고질적인 문제점은 사실성이 높은 고해상도의 물체를 렌더링 하는 데 있어서 방대한 양의 컴퓨팅 파워와, 시간 비용, 인력이 소모되는 것이다. 뉴럴 렌더링은 기존의 렌더링의 문제점을 개선하며 뛰어난 성능을 보여준다. 뉴럴 볼륨 렌더링은 장면으로 광선을 추적하고 광선의 길이에 대해 적분을 취해 이미지 또는 비디오를 생성하는 방법이다. 뉴럴 볼륨 렌더링의 대표적인 방식은 뉴럴 네트워크를 사용하여 암시적 표현을 정의하는 것이다. 많은 3D 이미지 생성 방식은 일반적으로 컨볼루션 아키텍처를 기반으로 하는 복셀, 메시, 포인트 클라우드를 사용한다. 뉴럴 렌더링에서 주목해야 할 기술 중 하나는 새로운 뷰 합성(View Synthesis)이다. 뷰 합성 문제에서 뉴럴 네트워크은 임의의 관점, 즉 한 번도 보지 못한 관점에서의 프레임을 렌더링 하는 방법을 학습한다.

뷰 합성을 위한 렌더링 과정은 현실세계와 밀접하다. 최근 실종자 수색, 또는 범죄자 몽타주를 3D 로 표현하여 장기 실종자를 찾는 데 성공한 사례가 있다. 또한 건축 분야에서도 다각도의 이미지를 렌더링 하여 집 내부를 구현할 수 있다. 이는 뉴럴 렌더링 과정으로 만든 3D 도면 기술에 사용되며, 3D 프린터로 집을 제작하여 빠른 생산이 가능한 3D 프린터 하우스에 사용될 수 있다. 대표적으로 자율주행에서 카메라로 찍힌 이미지들을 통해 자동차, 거리, 장애물 등 여러 요소들의 실제 거리를 측정할 수 있는 벡터 공간이 뉴럴 렌더링 과정을 통해 표현된다. 이처럼 2D 이미지를 사용해 3D 영상을 만드는 기술은 범죄, 자율주행, 모델링, 건축 등 다양한 분야에서 사용되며 이러한 3D 표현 네트워크들은 실제 사례에서 빈번하게 사용되는 만큼 악의적인 공격에 견고해야 한다.

따라서 본 논문에서는 적대적 공격의 위험성과 뉴럴 렌더링의 관련 연구에 대해 살펴본다.

* 교신저자 (corresponding author)

2. 배경

2.1 적대적 공격

적대적 사례(Adversarial Example)는 훈련 알고리즘의 근본적인 약점이라는 것을 시사한다. 이러한 적대적 사례의 원인은 불명확 했지만, 불충분한 모델 평균화와 전적인 지도 학습의 불충분한 정규화가 결합된 심층 뉴럴 네트워크의 극단적인 비선형성 때문이라고 추측되었다. 그러나 Ian Goodfellow 는 딥러닝 모델이 지나치게 선형적이며 결정 경계에 혼란을 주기 때문에 적대적 공격이 가능하다고 주장했다. 그가 제안한 FGSM(Fast Gradient Sign Method) [2]은 학습된 모델에 입력 값이 출력 값과 멀어지 도록 경사 하강법을 적용하여 기울기를 조작하도록 한다. 그 후 적대적 예제를 학습한 적대적 훈련이 드롭아웃만을 사용한 것보다 더 많은 정규화 이점을 제공할 수 있음을 보여준다. 선형성으로 인해 훈련하기 쉬운 모델을 설계하는 것과, 적대적 섭동(Adversarial perturbation)에 저항하기 위해 비선형 효과를 사용하는 모델을 설계하는 것 사이에 근본적인 텐션(tension)을 제안했다. 이후 PGD(Projected Gradient Descent) [3] 방식이 나왔다. 이것 또한 I-FGSM 방식의 응용이며, 차이점은 옵티머 델타 값 초기화 차이에 있다. PGD 는 랜덤한 포인트로 초기화를 하여 훨씬 빠르고 효과적이다. 현재 보편적으로 화이트박스 적대적 공격의 방법론으로 쓰인다.

UAP(Universal adversarial perturbations) [4]는 입력 데이터에 관계없이 뉴럴 네트워크에 혼선을 주는 보편적인 적대적 공격을 중심으로 연구를 진행했다. 그들은 데이터 포인트가 각자 클래스들의 결정 경계를 벗어나는 최소한의 벡터를 찾는 방식으로 노이즈를 생산했다. UAP 는 학습시간이 오래 걸리는 단점이 있어, 2019 년에 나온 UAP 의 아이디어를 한층 발전시킨 Fast-UAP [5]가 출간되었다. 이는 여러 벡터들 중에서 현재 노이즈 벡터와 코사인 유사도가 높은 벡터를 선택하여 학습하는 방식이다. UAP 에 비해서 오류가 높은 노이즈를 더 빨리 생성할 수 있다.

2.2 뉴럴 렌더링의 3D 표현

뉴럴 렌더링(Neural Rendering)은 고전적인 렌더링보다 완전 하고 현실적인 뷰로, 렌더링 하는 심층 뉴럴 네트워크과 기존의 3D 표현 및 렌더러를 결합한다. 뉴럴 네트워크를 이용한 3D 표현 방식에는 대표적으로 3 가지 방식이 있다.

암시적인 3D 표현 방법으로 NeRF [6]는 뷰 합성 태스크에서 3D 컨볼루션을 사용하지 않고 완전 연결 레이어만을 사용한 방법이다. 그들은 카메라를 이용하여 이미지를 찍은 과정에서 생기는 직선 Ray 에 위치한 모든 점들의 색상, 투명도의 합이 직선에 위치한 픽셀 값이라고 함수를 정의해 학습에 사용한다. 3D 위치 정보와 카메라의 위치 방향 등을 포함한 카메라 정보(viewing direction)을 입력으로 받아 색상과 투명도를 출력한다. 해당 연구는 완전 연결 레이어를 사용하기 때문에 3D 표현을 위한 많은 메모리 연산이 필요하지 않는 장점이 있다.

명시적인 3D 표현 방법으로 Neural Volumes [7] 연구가 있다. 그들은 다각도의 이미지들을 VAE(Variational Auto-Encoder)를 통해 특징 값들을 풍부하게 표현했고, 디코더 과정 중 특징 값을 3D 복셀로 만들어 3D 컨볼루션을 진행했다. 이들은 학습을 위해 많은 메모리를 사용하지만 3D 컨볼루션을 통해 NeRF 보다 비교적 낮은 연산량을 사용한다.

마지막으로 Chan 연구진들은 [8] 앞서 설명한 암시적 표현, 명시적 표현 방식의 장점들을 활용하여 Tri-plane hybrid 3D 표현을 제안한다. 그들은 StyleGAN2 생성기로 3 장의 특징 벡터들을 출력한다. 그 후 xy 평면, yz 평면, xz 평면에 맞게 특징 벡터들을 재배열한 후 평면들의 특징 값을 활용하여 색상, 투명도 정보를 출력한다. 이렇게 나타낸 3D 표현으로 뉴럴 렌더링을 가능하게 하며 다양한 문제에서 사용된다.

3D 표현 방식들로 나타낸 기하학적 3D 데이터는 고정된 이미지 세트들을 활용해 표현한다. 새로운 3D 표현 방법은 카메라 포즈에 따라 조정된 이미지 및 비디오 합성을 다룬다. 3D 표현의 주요 과제는 장면의 폐색 및 보이지 않는 부분을 유추하는 것이다. 컴퓨터 비전에서 이미지 기반 렌더링 방법은 일반적으로 장면 기하학을 새로운 뷰의 좌표 프레임으로 재구성하기 위해 최적화 기반 멀티 뷰 스테레오 방법에 의존한다.

3. 관련 연구

<표 1> 관련 연구 비교

	테스크	공격 방법	공격 목적
Wong [9]	스테레오 정합	FGSM	깊이 추정
Cheng [10]	스테레오 정합	PGD	깊이 추정
Berger [11]	스테레오 정합	UAP	깊이 추정
Wong [12]	Monocular	FGSM	깊이 추정
Mopuri [13]	Monocular	UAP	깊이 추정
Zhang [14]	3D 분류	DAI-FGSM	3D 표현

많은 연구자들은 딥러닝이 선형적이며 공격에 취약하다는 단점을 이용하여 다양한 연구를 진행하였고, 객체 분류와 탐지를 목표로 했다. 하지만 기술이 발전하면서 3D 표현, 스테레오 매칭 등 보다 복잡한 연구를 딥러닝과 접목시키기 시작했고 성능은 점점 발전하고 있다. 우리는 그 중 스테레오 매칭, Monocular, 3D 분류에 적대적 공격을 적용한 관련 연구들을 조사했다.

자율주행의 주요 기술인 벡터 공간을 추출하려면 스테레오 매칭을 통해 이미지의 깊이 맵을 알아야 한다. Wong 연구진들은 [9] 시간 축으로 이어진 2 개의 유사한 이미지들을 입력으로 주어 깊이 맵을 출력하는 네트워크에 보편적인 섭동 공격(adversarial perturbation attack)을 가하였다. 그들은 FGSM 방식을 사용하여 네트워크의 기울기 값이 실제 라벨과 반대

로 학습이 되도록 해 네트워크를 교란시켰다. Cheng 연구진들 [10] 은 스테레오 매칭 알고리즘을 이용한 네트워크를 더욱 견고 하게 만드는 연구를 진행했다. 그들은 projected gradient descent(PGD) 방식으로 학습한 적대적 공격 패치를 입력 이미지에 합성하여 스테레오 네트워크의 한계점을 알렸다. 그 후 더 견고한 네트워크를 제인 하는데, 컨볼루션 네트워크의 특징 값을 그대로 매칭하는 기존 연구들과 달리 그들은 센서스 변형을 적용하여 매칭하였다.

앞서 설명한 연구들은 각 입력 이미지 마다 다른 잡음을 적용했던 것과 달리 Berger 연구진들은[11] 스테레오 네트워크의 백본 네트워크와 데이터셋의 조합마다 보편적인 섭동을 생성하여 좀 더 보편적인 잡음을 생성했다. Wong 연구진들은 [12] 1 개의 이미지만을 입력으로 하여 단일 이미지의 깊이를 추정하는 단안 깊이 예측 네트워크(monocular depth prediction network)의 취약점을 연구했다. 그들은 경사 하강법을 사용해 예측 값과 실제 출력 값이 멀어지게 하는 잡음을 생성했다. Mopuri 연구진 [13] 또한 UAP 를 더 발전시킨 GD-UAP 알고리즘으로 패턴 노이즈 연구를 진행했다. 그들은 그동안 성능이 미미했던 블랙박스 상황에서 GD-UAP 을 통해 단일 깊이 추정뿐만 아니라 영상 분할과 인식이 가능하다고 설명했다.

Zhang 연구진들은 [14] 360-attack 이라고 불리는 구형의 3D VR 이미지에 적대적 공격을 가한 연구를 진행했다. 그들은 3D 이미지를 2D 평면 이미지 세트로 표현한 후 Distortion-Aware Iterative Fast Gradient Sign Method(DAI-FGSM) 기법을 사용한 공격을 가하여 교란시키는 효과를 입증했다.

4. 결론 및 기대효과

적대적 공격 기술이 발전 하면서 공격의 대상 또한 다양해지고 있다. 최근 각광받는 분야인 뉴럴 렌더링의 3D 표현에 적대적 공격이 가해질 경우 CG, 포렌식 범죄, 저작권 문제와 같이 일상의 밀접한 분야에 큰 혼란을 야기할 수 있다. 따라서 뉴럴 렌더링에 대한 적대적 공격 연구가 필수적이다. 본 논문은 해당 연구 분야에 도움이 될 것이라 기대한다.

사사

이 논문은 2022 년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No.2022R1A4A1033600). 또한, 본 연구는 2022 년 과학기술정보통신부 및 정보통신기획 평가원의 SW 중심대학사업의 연구결과로 수행되었음 (20180002160301001).

참고문헌

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

[2] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014

[3] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083, 2017.

[4] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversar- ial perturbations. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1765–1773, 2017.

[5] Jiazhu Dai and Le Shu. Fast-uap: An algorithm for expediting universal adversarial perturbation generation using the orientations of perturbation vectors. Neurocomputing, 422:109–117, 2021.

[6] Mildenhall, Ben, et al. "Nerf: Representing scenes as neural radiance fields for view synthesis." Communications of the ACM 65.1 (2021): 99-106.

[7] Lombardi, Stephen, et al. "Neural volumes: Learning dynamic renderable volumes from images." arXiv preprint arXiv:1906.07751 (2019).

[8] Chan, Eric R., et al. "Efficient geometry-aware 3D generative adversarial networks." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[9] Wong, Alex, Mukund Mundhra, and Stefano Soatto. "Stereopagnosia: Fooling stereo networks with adversarial perturbations." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35. No. 4. 2021.

[10] Kelvin Cheng, Christopher Healey, and Tianfu Wu. Towards adversarially robust and domain generalizable stereo matching by rethinking dnn feature backbones. arXiv preprint arXiv:2108.00335, 2021

[11] Zachary Berger, Parth Agrawal, Tian Yu Liu, Stefano Soatto, and Alex Wong. Stereoscopic universal perturbations across different architectures and datasets. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15180–15190, 2022

[12] lex Wong, Safa Cicek, and Stefano Soatto. Targeted adversarial perturbations for monocular depth prediction. Advances in neural information processing systems, 33:8486–8497, 2020

[13] Konda Reddy Mopuri, Aditya Ganeshan, and R Venkatesh Babu. Generalizable data-free objec- tive for crafting universal adversarial perturbations. IEEE transactions on pattern analysis and machine intelligence, 41(10):2452–2465, 2018.

[14] Yunjian Zhang, Yanwei Liu, Jinxia Liu, Jingbo Miao, Antonios Argyriou, Liming Wang, and Zhen Xu. 360-attack: Distortion-aware perturbations from perspective-views. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15035–15044, 2022.