Black-box 상황에서의 비디오 분류 모델 공격 연구 동향

박성빈, 이연준*
한양대학교 컴퓨터공학과 바이오인공지능융합전공, *한양대학교 컴퓨터공학과

pbt98@hanyang.ac.kr, *yeonjoonlee@hanyang.ac.kr

A Survey on Adversarial Attacks on Black-box Video Classifiers

Sung Bin Park, Yeon Joon Lee*
Major in Bio Artificial Intelligence, Department of Computer Science and Engineering, Hanyang University, * Department of Computer Science and Engineering, Hanyang University.

요 약

딥러닝 모델 기반 서비스들에 대한 적대적 기계학습 기법 기반의 공격은 늘어나는 추세다. 그 중, 비디오 분류 서비스에 대한 공격은 비용이 높아 이미지 분류 서비스보다 상대적으로 공격하기 어렵게 여겨졌다. 그러나, 최근 연구들은 이를 다양한 방법으로 해결하여 Black-box 상황에서 적은 비용으로 공격에 성공했다. 이에 본 논문은 Black-box 상황에서의 비디오 분류 서비스에 대한 적대적 기계학습 공격 기법들을 분류하고 앞으로의 연구 방향을 제시한다.

Ⅰ. 서 론

Youtube, Instagram, Facebook 과 같은 대형 비디오 플랫폼의 등장으로 비디오의 양은 절대적으로 증가하고 있다 [1]. 비디오의 절대적인 양이 증가함에 따라 정확하면서 속도가 빠른 비디오 분류 기법이 요구되어, 이를 충족하는 딥러닝 기반의 비디오 분류 기술이 활발히 연구되어왔다.

그러나, 딥러닝 기반의 기술들은 적대적 기계학습 공격 기법에 취약하다는 문제점을 가지고 있다. 그러나 본 기술의 경우, 2010 년대 중반까지 비디오의 차원이 높은 점, 긴 처리 시간, 그리고 많은 수의 비디오 분류 딥러닝 모델이 공개되어 있지 않은 문제 때문에 공격 비용이 높고 현실적인 시나리오 수립이 어려워 연구 진행이 더디었지만, 2010 년대 후반부터 공격 비용을 줄이고 Black-box 상황에서도 정확도가 높은 유의미한 적대적 기계학습 공격 연구들이 진행되기 시작했다.

비디오의 절대적인 양의 증가세와 이것이 사회에 미치는 파급력을 고려할 때, 비디오 분류 기술에 대한 공격은 위험성이 크고, 특히 Black-box 상황에서의 공격은 더욱 위험하므로, 본 논문에서는 현존하는 Black-box 상황에서의 비디오 모델에 대한 공격 연구들의 동향과 앞으로의 연구 방향에 대해 제시한다.

Ⅱ. 배경

1. 적대적 기계학습

적대적 기계학습은 딥러닝 모델의 결정 경계에 대한 정보를 얻거나 추측하여 이를 이용해 딥러닝 모델의 오작동을 일으킬 수 있는 적대적 샘플들을 생성하여 공격하는 기법이다.

적대적 샘플을 이용한 공격은 공격자의 의도 반영 가능 여부에 따라 Untargeted 공격과 Targeted 공격으로 나뉜다. Untargeted 공격은 딥러닝 모델의 오작동만을 일으킬 뿐, 공격자가 의도하는 대로 일으키지 못하는 공격이다. 반면, Targeted 공격은 공격자가 의도하는 대로 오작동을 일으키는 공격으로 Untargeted 공격보다 위험성이 높은 공격이다.

Ⅲ. 본론

본 논문에서는 Black-box 상황에서의 딥러닝 기반 비디오 분류 기술에 대한 적대적 기계학습 공격 연구의 동향을 알아본다. 공격 분류는 크게 두 가지로 나뉘는데, 첫번째는 Transferability 기반의 공격이고, 두번째는 Query 기반의 공격이다.

1. Transferability 기반 공격

Transferability 기반 공격은 딥러닝 파라미터가 대중에게 공개 되어있는 오픈소스 White-box 모델과, 공개되어있지 않은 Black-box 모델이 유사하다는 것을 이용한다. 즉, 공격자가 파라미터를 이미 알고 있는 White-box 딥러닝 모델에 대해서 공격을 진행한 다음, 여기서 얻어진 적대적 기계학습 파라미터들을 Black-box 딥러닝 모델 공격에 그대로 적용해 공격하는 방법이다.

[2] 는 White-box 이미지 모델로부터 Black-box 비디오 모델로의 Transferability 기반 공격을 성공했다. 비디오의 각각의 프레임에 해당하는 이미지를 인지하는 모델을 학습시켜 (White-box), 이 모델의 파라미터 정보를 이용해서 적대적 샘플 생성에 필요한 Perturbation 파라미터들을 찾는다. 그런 다음, 이를 Black-box 비디오 분류 모델에 적용시켜 적대적 샘플을 생성하고 Gradient 추측을 통해 샘플에 대한 미세 조정을 거쳐 공격을 성공했다.

비디오는 이미지와 달리 시계열적 특성을 지니는데, 여기서 딥러닝 모델은 각 프레임의 이미지의 정보 뿐만 아니라, 비디오의 시계열적 특성 또한 얻을 수 있어야 정확한 비디오 판단이 가능하다. 그러나, 시계열적 특성은 이미지 특성과 달리 딥러닝 모델마다 판단 결과가 매우 다

양하기 때문에, White-box 비디오 모델에서 Black-box 비디오 모델로의 Transferability 기반 공격이 어려웠다.

그러나, [3] 은 White-box 비디오 모델에서 Black-box 비디오 모델로의 Transferability 기반 공격을 성공했다. 이 연구는 비디오 모델 간의 시계열적 특성 인식의 편차를 줄여 Transferability 를 높이기 위해, 인접 프레임으로부터의 Gradient 의 정보를 함께 활용하여 적대적 샘플을 최적화하는 전략을 통해 6 개의 비디오 인식 모델에서 유효성을 입증했다.

2. Query 기반 공격

Query 기반 공격은 Gradient 를 Zero-order 최적화 기법을 통해 추측함으로써 딥러닝 모델의 결정 경계를 알아내고 적대적 샘플을 생성하는 공격이다. 이 공격에 쓰이는 최적화 기법에는 FD (Finite Differences) 나 NES (Natural Evolution Strategies) 가 쓰인다. 이 공격은 transferability 기반 공격보다 좋은 성능을 보이는 것이 특징이나, 공격 비용이 더 많이 들어가는 것이 단점이다. 이를 해결하기 위해 적대적 샘플을 생성하기 위한 Query 의 수를 줄여 공격 비용을 줄이는 방향으로 연구가 진행되어왔다.

[4] 에서는 중요도가 높은 프레임들의 중요한 부분에만 Perturbation 을 추가해 적대적 샘플을 생성함으로써 Query 의 수를 Untargeted 공격에 대해 기존 공격보다 28%를 줄였다. [4] 는 모든 프레임의 모든 부분을 사용하던 과거 연구와 달리, 중요 프레임을 추출하고, 비디오 인식 모델이 중요 프레임에서 집중적으로 봤던 부분들을 미리 알아내어 1 차적으로 Query 수를 줄이고, Heuristic 전략을 통해 2 차적으로 줄여 Untargeted 공격에서 Query 수를 기존보다 28% 줄였다.

[5] 에서는 중요 프레임 탐지 및 프레임 내 중요 부분 선택과 Perturbation 추가 과정이 독립적인 것에서 비롯되는 비효율성을 해결하여 더 정확한 중요 프레임 탐지와 적은 Perturbation 추가만으로 공격을 성공했다. [5] 는 Perturbation 추가 (행동), 비디오 인식 모델의 목표 레이블 확률 (보상) 을 기준으로 강화학습을 기반으로 공격을 설계하여, 비디오 인식 모델의 결과를 피드백 삼아 공격 모델이 [4] 보다 더 정교하게 Perturbation 을 추가할 수 있도록 만들었다.

한편, [6] 에서는 비디오에서는 모션이 있는 부분이 중요하다 여겨 모션이 있는 부분에 Perturbation 을 추가하여 Query 수를 줄이는 방법을 제시했다. 모션이 있는 부분은 비디오 분석 분야에서 자주 쓰이는 누적 모션 벡터 [7] 와 빛의 흐름 정보 [8] 를 생성하여 판단하였다. 이와 달리, [9] 에서는 아핀 변환과 같은 기하학적 변환을 사용해 Query 의 수를 줄였다. 프레임 내 물체의 기하학적 특성은 변화시키지 않으면서, 여러 위치로 평행이동, 회전 등의 처리를 거쳐 물체가 움직일 수 있는 방향을 알아내어 효율적인 적대적 샘플을 생성한다.

Ⅲ. 결론

본 논문에서는 Black-box 상황에서의 비디오 분류 모델에 대한 공격 연구의 동향에 대해 다루었다. 공격은 크게 Transferability 기반 공격과 Query 기반 공격이 주를 이루었다.

Transferability 기반 공격의 경우, 공격 성공률은 다소 떨어지지만 Query 기반 보다 공격 비용이 적은 장점이 있다. 그러나, 비디오의 시계열적 특성으로 인한 어려움 때문에 최근에 와서야 White-box 비디오 분류 모델에서 Black-box 비디오 분류 모델로의 Transferability 기반 공격이 처음 이루어졌다. 이를 미루어 보아, Transferability 기반 공격에서는 많은 비디오 분류 모델에 통용될 수 있는 적대적 샘플 생성 연구가 필요한 상황이다.

한편, Query 기반 공격의 경우, 공격 비용은 Transferability 기반 공격보다 높으나, 공격 성공률이 높은 장점이 있다. Black-box 상황에서는 무작위 기반 Gradient 추측 방법이 주가 될 수 밖에 없는데, 이를 그대로 모든 프레임에 적용하는 공격은 성공률이 낮을 뿐더러, Query 의 수도 많이 필요했다. 이에, 최근의 많은 연구들은 비디오에 대한 사전 지식 (모션이 있는 부분, 빛의 흐름, 중요 프레임 등) 을 활용하여 Perturbation 이 추가되어야 하는 부분을 제한함으로써 Query 의 수를 줄이고 공격 성공률을 높였다. 그러나, 이런 사전 지식들은 특정 비디오 종류에 대해서만 적용될 수 있어 현실적인 공격이 되기 어려운 경우들이 있기에 많은 비디오의 특성을 반영할 수 있는 사전 지식에 대한 발견이 필요하고, 사전 지식 외에도 정교한 적대적 샘플 생성을 위해 Perturbation 추가 범위를 줄일 수 있는 방법에 대한 연구가 필요하다.

ACKNOWLEDGMENT

이 논문은 2022 년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원 (NRF-2022R1F1A1074999) 을 받아 수행된 연구임.

참 고 문 헌

[1] Jason Wise, “How Many Videos Are on Youtube in 2023?”, (https://earthweb.com/how-many-videos-are-on-youtube/)

[2] Jiang, Linxi, et al. "Black-box adversarial attacks on video recognition models." Proceedings of the 27th ACM International Conference on Multimedia. 2019.

[3] Wei, Zhipeng, et al. "Boosting the Transferability of Video Adversarial Examples via Temporal Translation." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36. No. 3. 2022.

[4] Wei, Zhipeng, et al. "Heuristic black-box adversarial attacks on video recognition models." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 07. 2020.

[5] Wei, Xingxing, Huanqian Yan, and Bo Li. "Sparse black-box video attack with reinforcement learning." International Journal of Computer Vision 130.6 (2022): 1459-1473.

[6] Zhang, Hu, et al. "Motion-excited sampler: Video adversarial attack with sparked prior." European Conference on Computer Vision. Springer, Cham, 2020.

[7] Wu, C.Y., Zaheer, M., Hu, H., Manmatha, R., Smola, A.J., Kra henbu hl, P.: Com- pressed video action recognition. In: Proceedings of the IEEE Conference on Com- puter Vision and Pattern Recognition. pp. 6026- 6035 (2018)

[8] Chambolle, A.: An algorithm for total variation minimization and applications. Journal of Mathematical imaging and vision 20(1-2), 89- 97 (2004)

[9] Li, Shasha, et al. "Adversarial attacks on black box video classifiers: Leveraging the power of geometric transformations." Advances in Neural Information Processing Systems 34 (2021): 2085-2096.