한국컴퓨터정보학회 하계학술대회 논문집 제31권 제2호 (2023. 7)

적대적 머신러닝 공격과 방어기법

이제민*, 박재경O
*한국폴리텍대학 강서캠퍼스,
O한국폴리텍대학 강서캠퍼스
e-mail: jm98@daum.net*, jakypark@kopo.ac.krO

A Study Adversarial machine learning attacks and defenses

jemin Lee*, Jae-Kyung ParkO
*Dept. Korea Polytechnics Gangseo Campus,
ODept. Korea Polytechnics Gangseo Campus

● 요 약 ●

본 논문에서는 기계 학습 모델의 취약점과 대응책에 초점을 맞추어 적대적인 기계 학습 공격 및 방어 분야를 탐구한다. 신중하게 만들어진 입력 데이터를 도입하여 기계 학습 모델을 속이거나 조작하는 것을 목표로 하는 적대적 공격에 대한 심층 분석을 제공한다. 이 논문은 회피 및 독성 공격을 포함한 다양한 유형의 적대적 공격을 조사하고 기계 학습 시스템의 안정성과 보안에 대한 잠재적 영향을 조사한다. 또한 적대적 공격에 대한 기계 학습 모델의 견고성을 향상시키기 위해 다양한 방어 메커니즘과 전략을 제안하고 평가한다. 본 논문은 광범위한 실험과 분석을 통해 적대적 기계 학습에 대한 이해에 기여하고 효과적인 방어 기술에 대한 통찰력을 제공하는 것을 목표로 한다.

키워드: 적대적 기계 학습(adversarial machine learning), 취약성(vulnerability), 회피 공격(evasion attack), 중독 공격 (poisoning attack)

I. Introduction

적대적 머신러닝은 최근 인공지능 기술의 발전으로 인해 중요성이 증가하고 있는 분야다. 이 기술은 머신러닝 모델을 공격하거나 속이는 공격 기법을 의미한다. 예를 들어, 적대적 샘플을 생성하여 모델의 오분류를 유도하거나, 적대적인 조작을 통해 모델의 동작을 왜곡시킬 수 있다. 이러한 적대적 공격은 인공지능 시스템의 신뢰성과 안전성을 위협할 수 있다. 따라서, 적대적 머신러닝 공격에 대한 방어 기법의 개발이 중요한 연구 주제로 부각되고 있다. 이 논문은 적대적 머신러닝 공격과 방어에 대한 현재의 이해와 동향을 조사하고, 새로운 방어 기법의 개발과 평가를 목표로 한다. 이를 통해 보다 안전하고 신뢰할 수 있는 인공지능 모델을 구축하는 데에 기여하고자 한다.

II. Preliminaries

1. 적대적 공격 유형

1-1. 회피공격 :

적대적 사례라고도 하는 회피 공격은 추론 중에 기계 학습 모델을 속이기 위해 입력 데이터를 조작하는 것과 관련된다. 목표는 입력에 눈에 띄지 않는 섭동을 추가하여 모델을 잘못 분류하거나 잘못된 결과를 생성하는 것이다. 그래디언트 기반 방법과 같은 다양한 최적화 기술을 사용하여 이러한 적대적인 예를 생성한다.

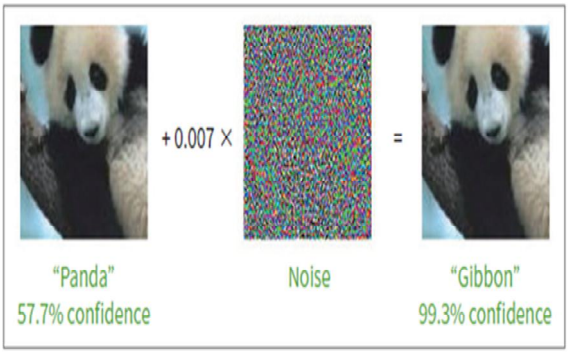

Fig. 1. Ian Goodfellow, Joonathan Shlens, and Christian Szegedy, Explaining and harnessing adversarial examples, 2014

1-2. 중독 공격 : 중독 공격은 교육 데이터 세트에 악의적이거나 조작된 데이터를 주입하여 모델의 교육 프로세스를 손상시키는 것을 목표로 한다. 이러한 오염된 샘플을 도입함으로써 공격자는 모델의 학습된 결정 경계를 편향시키거나 전체 성능을 저하시키려고 한다. 중독 공격은 손상된 모델이 학습 후에도 계속해서 잘못된 예측을 하므로 오래 지속되는 영향을 미칠 수 있다.

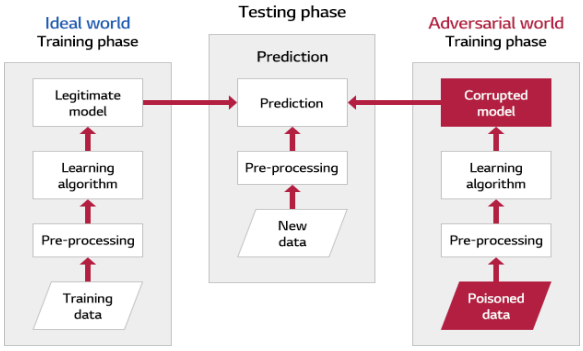

Fig. 2. Manipulating Machine Learning: Poisoning attacks and Countermeasures for Regression Learning

1-3. 공격 생성 기술 및 알고리즘

1. FGSM(Fast Gradient Sign Method): 이 기법은 모델의 예측 오류를 최대화하기 위해 손실 함수의 기울기 방향으로 입력 특징을 교란시킨다.

2. PGD(Projected Gradient Descent): 미리 정의된 최대 섭동 예산 내에서 입력에 작은 섭동을 반복적으로 적용하여 FGSM을 확장하여 적대적인 예가 인식 가능한 범위 내에 있도록 한다.

3. JSMA(Jacobian-based Saliency Map Attack): 이 방법은 가장 영향력 있는 기능을 식별하고 변경 사항을 최소화하면서 오분류를 유발하도록 수정한다.

4. 데이터 포이즈닝 공격: 데이터 주입, 데이터 수정 또는 모델 백도어링과 같은 기술을 사용하여 훈련 데이터 세트를 조작하고 모델의 학습된 매개변수를 편향시킨다.

2. 적대적 방어 메커니즘

2-1. 회피 공격에 대한 방어

1. 적대적 훈련(Adversarial Training): 이 방법은 훈련 과정에서 생성된 적대적 사례로 훈련 데이터를 보강하는 것이다. 모델을 이러한 적대적 사례에 노출함으로써 모델은 더 견고해지고 보이지 않는 적대적 입력에 대해 더 잘 일반화된다.

2. 방어적 증류: 방어적 증류에는 더 복잡한 모델 또는 앙상블 모델의 동작을 모방하도록 모델을 교육하는 것이 포함된다. 부드러운 확률을 생성하도록 모델을 훈련하면 적대적인 섭동에 더 탄력적으로 된다.

3. 그라디언트 마스킹: 이 방어 기술은 모델의 그라디언트가 공격자에게 덜 유용하도록 만드는 것을 목표로 한다. 여기에는 역전파 프로세스 중에 임의 노이즈를 추가하거나 그래디언트 정보를 난독화하여 공격자가 효과적인 섭동을 찾기 어렵게 만든다.

4. 무작위화: 무작위화 기술은 모델이나 입력 데이터에 무작위성을 도입하여 공격자가 성공적인 적대적 섭동을 찾기 어렵게 만든다. 여기에는 입력에 노이즈를 추가하거나 추론 중에 무작위 변환을 적용하는 것이 포함될 수 있다.

2-2. 중독 공격에 대한 방어

1. 데이터 삭제: 데이터 삭제 기술은 교육 데이터 세트에서 악의적이거나 중독된 데이터를 탐지하고 제거하는 것을 목표로 한다. 여기에는 의심스럽거나 악의적인 샘플을 식별하고 폐기하기 위한 이상값 감지, 이상 감지 또는 데이터 필터링 기술이 포함될 수 있다.

2. 강력한 훈련 알고리즘: 강력한 훈련 알고리즘은 오염된 데이터에 대한 모델의 탄력성을 높이도록 설계되었다. 여기에는 훈련 프로세스 중에 적대적이거나 중독된 샘플의 존재를 명시적으로 고려하는 강력한 최적화와 같은 기술이 포함된다.

3. 입력 유효성 검사 및 필터링: 이 방어 메커니즘에는 추론 단계에서 입력 데이터의 무결성과 신뢰성을 확인하는 것이 포함된다. 의심스럽거나 악의적인 동작을 나타내는 입력을 식별하고 거부하기 위한 이상 탐지 또는 입력 필터링과 같은 방법이 포함된다.

4. 보안 집계: 보안 집계 기술은 협업 기계 학습 설정에서 집계된 데이터의 무결성을 보호하는 것을 목표로 한다. 보안 프로토콜과 암호화 기술을 사용하여 공격자가 집계 프로세스에 중독된 데이터를 주입하는 것을 방지한다.

III. The Proposed Scheme

적대적 기계 학습의 미래 방향과 열린 과제

1. 적대적 기계 학습의 새로운 트렌드: 적대적 기계 학습 분야가 계속 발전함에 따라 새로운 트렌드와 개발에 대한 최신 정보를 유지하는 것이 중요하다. 강화 학습, 생성 모델 또는 그래프 신경망과 같은 새로운 도메인에서 적대적 공격 및 방어 탐색과 같은 새로운 트렌드를 살펴봐야한다. 또한 블랙박스 공격 또는 물리적 세계 공격과 같은 새로운 공격 기술의 채택과 이러한 지능형 적대적 위협에 대응하기

위한 방어 메커니즘 개발을 할 필요가 있다.

2. 미개척 연구 분야 및 기회: 적대적 기계 학습에서 상당한 진전이 이루어졌지만 여전히 많은 미개척 연구 영역과 미개척 기회가 있다. 서로 다른 모델 및 도메인에 걸친 적대적 공격 및 방어의 이전 가능성 연구, 다중 모달 데이터에 대한 적대적 공격의 영향 조사 또는 방어 메커니즘의 해석 가능성을 향상시키기 위해 설명 가능한 AI 기술의 사용 탐색과 같은 잠재적인 연구를 해야한다.

3. 한계 극복 및 수비 전략 개선: 적대적 기계 학습 방어 기술은 몇 가지 제한 사항과 과제에 직면해 있다. 적응형 공격에 대한 방어의 취약성 또는 일부 방어 메커니즘과 관련된 높은 계산 비용과 같은 이러한 제한 사항에 대해 설명한다. 지속적인 학습을 통해 적응형 공격에 대한 강력한 방어를 개발하거나 대규모 데이터 세트 및 실시간 애플리케이션을 처리할 수 있는 효율적이고 확장 가능한 방어 기술을 활용하는 등 이러한 한계를 극복하기 위한 전략을 탐구해야한다.

4. 미래 R&D를 위한 제안: 적대적 기계 학습 분야의 향후 연구 및 개발에 대한 제안을 제공한다. 이는 기계 학습, 사이버 보안 및 인지 과학의 전문가를 포함하여 적대적 공격으로 인한 다각적인 문제를 해결하기 위한 학제 간 협력의 필요성을 강조한다. 또한 서로 다른 방어 기술 간에 공정하고 재현 가능한 비교를 가능하게 하기 위해 표준화된 평가 프레임워크, 데이터 세트 및 메트릭을 개발하는 것의 중요성을 강조한다. 또한 강력하고 탄력적인 기계 학습 모델의 설계를 위해 게임 이론 또는 진화 알고리즘을 활용하는 것과 같은 혁신적인 접근 방식을 탐색할 것을 제안한다.

Ⅳ. Conclusions

결론적으로, 이 논문은 적대적인 위협에 직면한 기계 학습 시스템의 보안 및 견고성을 향상시키는 것을 목표로 적대적인 기계 학습 공격 및 방어에 대한 포괄적인 탐색을 제공했다. 연구 전반에 걸쳐 회피 및 포이즈닝 공격을 포함한 다양한 유형의 적대적 공격을 조사하고 머신 러닝 모델의 무결성 및 신뢰성에 대한 잠재적 영향을 분석했다.

또한 향후 방향과 열린 과제를 제시하여 새로운 트렌드, 미개척 연구 분야 및 추가 발전 기회를 식별했다. 기존 방어 기술의 한계를 극복하고 방어 전략을 개선하는 것은 적의 공격에 한발 앞서가는 데 중요하다. 공동 노력, 학제 간 연구 및 표준화된 평가 프레임워크의 개발은 해당 분야의 발전을 촉진하는 데 필수적이다.

이 논문에 제시된 연구는 적대적 기계 학습에 대한 지식의 증가에 기여한다. 적대적 공격의 특성을 이해하고, 방어 메커니즘을 평가하고, 윤리적 및 법적 의미를 고려함으로써 우리는 적대적 위협을 견딜 수 있는 보다 안전하고 강력한 기계 학습 시스템을 구축하기 위해 노력할 수 있다.

이 분야가 계속 발전함에 따라 연구원, 실무자 및 정책 입안자가 효과적인 방어 전략을 개발하고 구현하는 데 경계하고 능동적으로 유지하는 것이 중요하다. 그렇게 함으로써 우리는 적대적 공격과 관련된 위험을 완화하고 컴퓨터 비전, 자연어 처리 및 사이버 보안을 포함한 다양한 영역에서 기계 학습 애플리케이션의 신뢰성과 신뢰성을 보장할 수 있다.

궁극적으로 목표는 적대적 공격이 완화되고 이러한 기술의 이점이 완전히 실현될 수 있는 기계 학습을 위한 탄력적이고 안전한 생태계를 구축하는 것이다. 지속적인 연구, 협업 및 책임 있는 AI 관행을 통해 기계 학습 시스템이 적대적 위협을 효과적으로 방어하고 민감한 데이터를 보호하며 다양한 산업에서 AI 기반 솔루션의 광범위한 채택을 가능하게 하는 미래를 위한 길을 열 수 있다.

우리는 이 복잡하고 진화하는 분야의 다양한 측면을 다루는 적대적 기계 학습 공격 및 방어에 대한 포괄적인 연구를 제시했다. 이 연구를 통해 귀중한 통찰력을 제공하고 효과적인 방어 전략을 제안했으며 향후 발전을 위한 주요 과제와 기회를 식별했다. 이 작업이 안전하고 탄력적인 기계 학습 시스템을 구축하기 위한 지속적인 노력에 기여하여 궁극적으로 우리 사회에서 AI 기술의 신뢰와 신뢰성을 높이는 것이 우리의 희망이다.

REFERENCES

[1] Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In IEEE Symposium on Security and Privacy (SP) (pp. 39-57). IEEE.

[2] Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., ... & Roli, F. (2017). Evasion attacks against machine learning at test time. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 387-402). Springer.

[3] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

[4] Huang, L., Joseph, A. D., Nelson, B., Rubinstein, B. I., & Tygar, J. D. (2011). Adversarial machine learning. In Proceedings of the 4th ACM workshop on Security and artificial intelligence (pp. 43-58).

[5] Liu, Y., Chen, X., Liu, C., & Song, D. (2017). Delving into transferable adversarial examples and black-box attacks. In International Conference on Learning Representations (ICLR).

[6] Athalye, A., Carlini, N., & Wagner, D. (2018). Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In International Conference on Machine Learning (ICML) (pp. 274-283).