신경망 내 가상 미끼 뉴런 배치를 통한 모델 인버전 공격 방어 매커니즘 연구

이선우[1], 노현우[2], 김수하[1], 박민서[1], 최우현[1], 윤승현[1]

[1]한국에너지공과대학교, [2]전남과학고등학교

{sunwoolee, water03, westpark, woohyunchoi, syoon}@kentech.ac.kr

jshs251208@h.jne.go.kr

A Study on Defense Mechanisms Against Model Inversion Attacks via Virtual Decoy Neuron Deployment in Neural Networks

Sunwoo Lee[1], Hyeonwoo Roh[2], Suha Kim[1], Minseo Park[1], Woo-Hyun Choi[1], and Seunghyun Yoon[1]

[1]Korea Institute of Energy Technology (KENTECH)

[2]Jeonnam Science High School

요 약

공격자가 모델 파라미터와 구조에 접근 가능한 환경에서, 모델 인버전 공격은 입력 x를 최적화 변수로 두고 목표 클래스 y에 대한 손실을 반복적으로 최소화함으로써 특정 클래스의 대표 입력을 복원할 수 있으며, 이때 입력 그래디언트 ∇x의 방향이 복원 품질에 중요한 영향을 미친다. 본 연구는 딥러닝 모델의 은닉층에 가상 미끼 뉴런(Virtual Decoy Neurons, VDN)을 삽입하여 공격 최적화 과정에서의 그래디언트 구성을 교란하는 방어 매커니즘을 제안한다. VDN은 기본 분류 경로와 분리된 미끼 경로 및 게이트로 구성되며, 정상 입력에서는 게이트를 억제해 성능 영향을 최소화하는 반면, 공격 상황에서는 미끼 경로가 입력 그래디언트 형성에 지배적으로 기여하도록 학습된다. 이를 위해 정상 데이터에서의 비활성화 항, 미끼 입력 분포에서의 활성화 항, 그리고 그래디언트 지배성을 유도하는 제약 항을 포함한 목적함수를 설계하였다. 제안 기법은 공격 최적화가 실제 데이터 분포에 기반한 복원으로 수렴하는 것을 어렵게 하고, 방어자가 의도한 미끼 목표 방향으로의 편향을 유도하는 것을 목표로 한다.

Ⅰ. 서 론

모델 인버전 공격(Model Inversion Attack)은 모델에 내재화된 정보를 악용해 특정 클래스의 대표 입력 또는 학습 데이터의 민감한 속성을 복원하는 공격이다. 특히 공격자가 모델 파라미터와 구조에 접근 가능한 환경(예: 모델 유출, 내부자 위협, 또는 공개 모델의 화이트박스 접근)에서는, 입력 x를 최적화 변수로 두고 목표 클래스 y에 대한 손실을 반복적으로 최소화함으로써 대표 입력을 생성할 수 있으며, 이 과정에서 입력 그래디언트 ∇x 가 복원의 핵심 신호로 작동한다[1],[2]. 기존 방어인 차등 프라이버시(DP)나 정보 제한 기반 기법은 정확도 저하, 학습 비용 증가, 또는 생성 기반/최적화 기반 공격에 대한 불완전한 견고성 등 한계를 가진다[2],[3]. 이에 본 연구는 은닉층에 가상 미끼 뉴런(Virtual Decoy Neurons, VDN)을 삽입하여 정상 추론에서는 미끼 경로를 억제해 성능 영향을 최소화하는 반면, 공격 최적화 과정에서는 미끼 경로가 입력 그래디언트 형성에 상대적으로 더 크게 기여하도록 학습함으로써 공격의 최적화 방향을 방어자가 설정한 미끼 목표로 편향시키는 방어 메커니즘을 제안한다.

Ⅱ. 위협 모델

그림 1과 같이 공격자는 대상 모델의 파라미터 θ에 접근할 수 있으며, 모델 구조와 손실 함수 형태를 알고 있는 화이트박스 조건에서 모델 인버전 공격을 수행한다고 가정한다. 공격자는 학습에 사용된 원시 데이터에 직접 접근하지 못하지만, 모델이 특정 클래스 y에 대해 높은 신뢰도를 출력하도록 입력 x를 반복적으로 최적화함으로써 해당 클래스의 대표 입력을 복원하거나, 복원된 입력으로부터 학습 데이터의 민감한 속성을 추정하는 것을 목표로 한다. 공격 과정은 입력 x를 최적화 변수로 두고 분류 손실과 입력 정규화 항을 최소화하는 형태로 정식화되며(식 1), 일반적으로 입력 그래디언트 ∇x를 이용한 반복 갱신으로 수행된다(식 2).

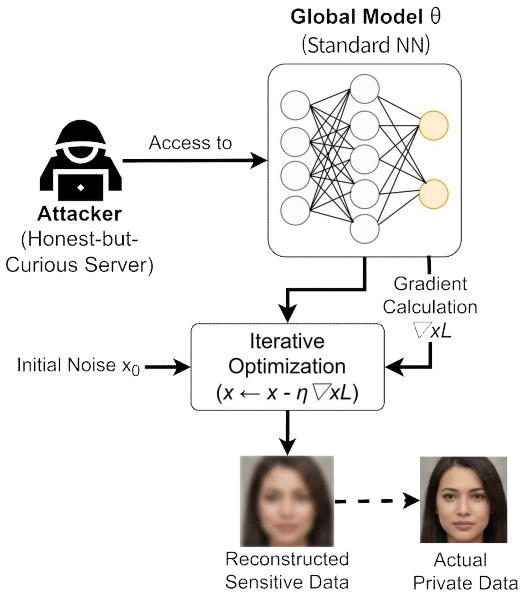

그림 1. 위협 모델 (모델 인버전 공격)

이를 위해 x를 최적화 변수로 두고 목적함수(식 1)를 최소화한다[2].

$$L_{atk}(x;\theta,y) = L(f(x;\theta),y) + \lambda R(x) \qquad \text{(식 1)}$$

여기서 $f(x;\theta)$는 대상 모델, L_{ce}는 분류 손실, $R(x)$는 입력의 자연스러움을 유도하는 정규화 항이며 λ는 정규화 강도이다. 공격은 식 2와 같이 일

반적으로 그래디언트 기반 반복 갱신으로 수행된다[1][2].

$$x^{(t+1)} = \Pi_x(x^{(t)} - \eta \Delta_x L_{atk}(x^{(t)}; \theta, y)) \quad \text{(식 2)}$$

η는 학습률, Π_x는 입력 도메인 x로의 프로젝션 연산이다. 반복이 진행되면 $x^{(T)}$는 목표 클래스 y에 대해 모델이 강한 신뢰도를 갖도록 갱신되며, 그 과정에서 학습 데이터 분포의 특징이 입력에 반영되어 프라이버시 침해로 이어질 수 있다.

Ⅲ. 제안하는 기법

가상 미끼 뉴런은 그림 2와 같이 은닉층에 추가되는 미끼 경로로서, 정상 입력에서는 거의 영향을 주지 않지만 공격자가 입력을 최적화할 때에는 입력 그래디언트가 미끼 목표 방향으로 형성되도록 유도한다.

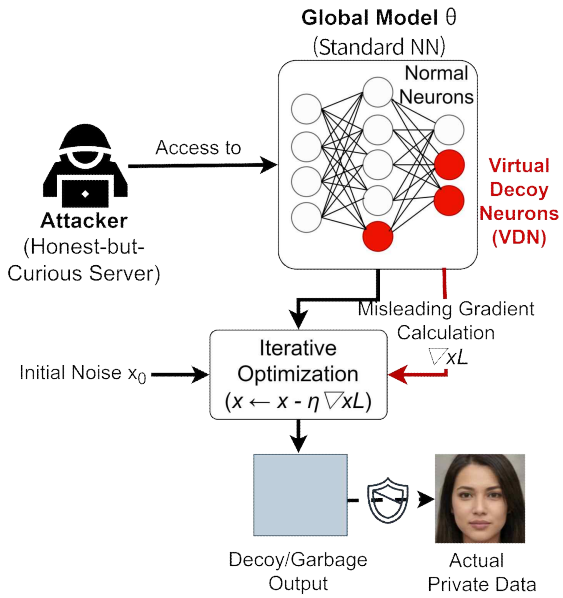

그림 2. VDN의 방어 원리

이를 위해 기본 분류 경로와 분리된 미끼 경로를 구성하고, 미끼 경로의 기여도를 입력에 따라 조절하는 게이트를 함께 학습한다. 입력 x에 대해 은닉 표현을 $h(x) = \phi(x; \theta_h)$라 두고, 기존 분류 경로의 logits를 $z_{task}(x) = g(h(x); \theta_t)$로 정의한다. VDN은 동일한 은닉표현 $h(x)$로부터 미끼 logits $z_{decoy}(x) = d(h(x); \theta_d)$를 생성하고, 게이트 $s(x) = \sigma(v(h(x); \theta_v)) \in [0,1]^m$를 통해 미끼 경로가 최종 출력에 기여하는 정도를 제어한다. 최종 logits는 식 3과 같이 결합한다.

$$z(x) = z_{task}(x) + \alpha(s(x) \odot z_{decoy}(x)), \quad \text{(식 3)}$$
$$f(x; \theta) = softmax(z(x))$$

여기서 ⊙는 원소별 곱이며, α는 미끼 경로의 영향도를 조절하는 상수이다. 정상 추론에서 분류 성능을 유지하려면 대부분의 정상 입력에 대해 $s(x) \approx 0$이 되도록 학습되어야 한다. 반대로 공격 상황에서는 입력 갱신에 사용되는 Δ_x가 미끼 경로의 영향을 받도록 설계되어, 공격 최적화가 실제 데이터 중심이 아니라 미끼 목표로 유도되도록 한다. 이를 만족시키기 위해 학습 손실을 식 4와 같이 구성한다.

$$L_{total} = L_{task} + \beta L_{suppress} + \gamma L_{amplify} + \delta L_{grad} \quad \text{(식 4)}$$

L_{task}는 원래의 분류 성능을 위한 교차엔트로피 손실이다. $L_{suppress}$는 식 5와 같이 정상 데이터에서 VDN 게이트가 켜지는 것을 억제하여 추론 성능 저하를 막는다.

$$L_{suppress} = E_{(x,y) \sim D}[\|s(x)\|_1] \quad \text{(식 5)}$$

또한 방어자가 정의한 미끼 입력 분포 D_{decoy}에 대해서는 식 6과 같이 게이트가 활성화되도록 한다.

$$L_{amplify} = E_{x \sim D_{decoy}}[|1 - s(x)|_1] \quad \text{(식 6)}$$

마지막으로 L_{grad}는 공격자가 입력을 최적화할 때 미끼 경로가 입력 그래디언트 형성에 더 크게 기여하도록 식 7과 같이 유도한다. 구체적으로, 미끼 입력 영역에서 미끼 경로의 입력 그래디언트 노름이 task 경로보다 작아지지 않도록 마진 $\kappa > 1$를 두어 다음과 같은 제약으로 학습한다.

$$L_{grad} = E_{x \sim D_{decoy}}[\max(0, \kappa \|\nabla_x L_{task}(x)\|_2 - \|\nabla_x L_{decoy}(x)\|_2)] \quad \text{(식 7)}$$

여기서 $L_{decoy}(x)$는 미끼 경로가 특정 미끼 목표로 향하도록 정의된 보조 손실이다. 위 손실 구성으로 인해 정상 입력에서는 $s(x)$가 억제되어 분류 성능이 유지되고, 공격자가 입력을 갱신하는 구간에서는 미끼 경로가 만들어내는 그래디언트 성분이 상대적으로 커져 공격 최적화 경로가 미끼 목표로 기울어지도록 유도된다.

Ⅳ. 결론

가상 미끼 뉴런(VDN) 기반 방어를 제안하였다. 공격자가 모델 파라미터 및 구조에 접근 가능한 환경에서 제안하는 VDN 기법은 공격 시 입력 그래디언트의 방향을 교란해 복원을 어렵게 하며, 정상 추론에서는 미끼 경로를 억제해 성능 영향을 최소화한다. 다만 적응형 공격자가 미끼 구조를 파악하거나 제거하면 방어 효과가 저하될 수 있다.

ACKNOWLEDGMENT

본 논문은 과학기술정보통신부 정보통신기획평가원에서 지원한 국산 AI 반도체 기반 마이크로 데이터센터 운영 및 확산 기술 개발 과제로 수행된 연구임 (과제번호: RS-2025-25457382)

참 고 문 헌

[1] Fredrikson, M. *et al.*, "Model Inversion Attacks that Exploit Confidence Scores and Basic Countermeasures," *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 2015.

[2] Zhang, Y. *et al.*, "The Secret Revealer: Generative Model Inversion Attacks Against Deep Neural Networks," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[3] Abadi, M. *et al.*, "Deep Learning with Differential Privacy," *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2016.