신뢰할수있는안공지능

▼ Assignment #2

DeepXplore와 FGSM/PGD 공격 기법의 수학적 결합 가능성에 대한 제안

DeepXplore는 coverage-guided fuzz testing 원리를 deep learning에 적용한 도구로, neuron coverage maximization을 목표로 입력을 생성합니다.

주어진 입력 x와 두 모델 f1,f2가 있을 때, DeepXplore는 다음과 같은 objective를 최적화합니다:

$\max_\delta \left[\lambda_1 \cdot D(f_1(x+\delta), f_2(x+\delta)) + \lambda_2 \cdot C(x+\delta)\right]$

- D(·,·): discrepancy를 측정 (softmax output의 L1 또는 L2 distance 등)
- C(x): neuron coverage 함수
- λ1,λ2: 각 항의 weight를 조절하는 하이퍼파라미터

이 목적함수는 gradient ascent 방식으로 최적화되지만, coverage term이 binary threshold 기반이거나 discrete일 수 있어 local minimum에 빠지기 쉽고 탐색의 효율성이 떨어질 수 있습니다.

FGSM/PGD와의 결합 아이디어

이러한 한계를 보완하기 위해, FGSM 및 PGD를 DeepXplore의 gradient 기반 테스트와 결합할 수 있습니다.

FGSM: $x' = x + \varepsilon \cdot \text{sign}(\nabla_x J(f(x), y))$

PGD: $x^{(t+1)} = \Pi_{B_\varepsilon(x)} \left[x^{(t)} + \alpha \cdot \text{sign}(\nabla_x J(f(x^{(t)}), y))\right]$

- Π: ε-ball 내로 projection
- J: loss function (기본적으로 cross-entropy지만 확장 가능)

Loss Function의 확장: Coverage-Aware 공격

기존 loss를 다음과 같이 확장할 수 있습니다:

$J(x, y) = \lambda_1 \cdot \text{CrossEntropy}(f(x), y) + \lambda_2 \cdot (1 - C(x))$

또는, PGD의 목적함수를 DeepXplore와 유사하게 정의:

$$x^{(t+1)} = \arg\max_{\|x - x^{(0)}\|_\infty \leq \varepsilon} [D(f_1(x), f_2(x)) + \gamma \cdot C(x)]$$

이러한 방식은 adversarial attack을 단순한 misclassification에서 확장하여, coverage maximization과 differential behavior detection을 동시에 달성하는 방향으로 해석될 수 있습니다.

DeepXplore의 objective와 FGSM/PGD의 gradient 기반 perturbation 전략은 수학적으로 정합성이 높으며, 서로 보완적으로 작동할 수 있습니다. 특히 coverage function을 differentiable surrogate로 근사할 경우, 더 효율적이고 강력한 test case generation이 가능해집니다.