DeepXplore와 FGSM/PGD 공격 기법의 결합 가능성에 대한 제안

DeepXplore는 neuron coverage를 기반으로 다양한 neuron activation pattern을 유도하여, neural network의 동작 범위를 탐색하고 잠재적인 오류 사례를 발견하는 testing framework이다. 반면 FGSM(Fast Gradient Sign Method)과 PGD(Projected Gradient Descent)는 neural network의 입력에 미세한 perturbation을 가해 prediction을 교란하는 대표적인 adversarial attack 기법이다. 이 두 기술은 목적은 다르지만 공통적으로 gradient 정보를 활용하여 모델의 취약 지점을 노출시킨다는 점에서 유사한 부분을 확인할 수 있다.

DeepXplore의 핵심은 model 간 output discrepancy를 유도하면서도 neuron coverage를 최대화하는 입력을 생성하는 데 있다. 하지만 gradient를 사용하는 input generation은 local minimum에 빠지거나, 계산 속도 문제로 인해 충분히 다양한 test case를 확보하지 못하는 문제가 있다. 이 지점에서 FGSM과 PGD 기법을 보조적으로 활용하면 다음과 같은 이점이 있다.

첫째, DeepXplore가 생성한 초기 input에 FGSM이나 PGD를 적용함으로써, 추가적인 decision boundary 근처의 미세한 perturbation을 유도할 수 있다. 이는 model의 약한 지점을 더 정밀하게 탐색하는 데 도움을 준다. 특히 PGD는 iterative 방식으로 더 강력한 attack을 수행하므로, test input을 더욱 다양화할 수 있을 것이다.

둘째, FGSM이나 PGD로 생성한 adversarial example을 DeepXplore의 differential testing 대상 model들에 적용함으로써, 단순한 misclassification이 아닌 output discrepancy를 유도하는 input으로 사용할 수 있다. 이로 인해 기존 DeepXplore보다 더 빠르게 suspicious behavior를 탐색할 수 있을 것이다.

셋째, DeepXplore의 neuron coverage 정보를 기반으로 한 weighted PGD 공격을 설계할 수 있다. 예를 들어 neuron coverage가 낮은 방향으로 PGD를 유도하거나, 이전에 활성되지 않았던 neuron을 최대한 자극하는 방향으로 attack을 수행할 수 있다. 이는 단순한 예측 오류 탐지를 넘어 coverage-guided adversarial testing이라는 확장된 개념을 제시한다.

물론 이러한 결합은 test reliability 측면에서 신중한 설계가 필요하다. FGSM과 PGD는 때때로 자연스러운 입력 분포에서 벗어난 예제를 생성할 수 있으므로, 실제 오류를 반영하는지 확인하는 후처리 절차가 필요하다.

결론적으로 FGSM과 PGD는 DeepXplore의 한계를 보완하면서 더 정교한 testing과 취약점 탐지를 가능하게 하는 도구가 될 수 있다. 이는 테스트의 공격성과 coverage 측면 모두에서 이점을 제공하며, 신뢰 가능한 neural network 개발을 위한 다음 단계로 확장될 수 있는 유의미한 접근이다.