고려대학교 KOREA UNIVERSITY

윤지원 교수 (고려대학교 정보보호대학원)

딥러닝 기반의 적대적 공격 탐지 장치 및 방법
(기술분류-보안-디지털 취약점 분석·대응)

기술성 분석

기술 개요

- 본 기술은 딥러닝 모델 또는 기계학습 모델을 대상으로 하는 적대적 공격을 탐지하는 장치 및 방법에 관한 것으로, 기본 딥러닝 모델이 탐지 기능을 수행하도록 함으로써 모델 구조를 달리하거나 알고리즘을 달리하는 등의 노력 없이 적대적 공격을 탐지할 수 있음
- 또한, 모델 업데이트가 필요한 경우에 기존의 기본 모델만 업데이트하여도 되며, 공격 탐지 모델을 보관할 별도의 공간을 마련할 필요가 없음

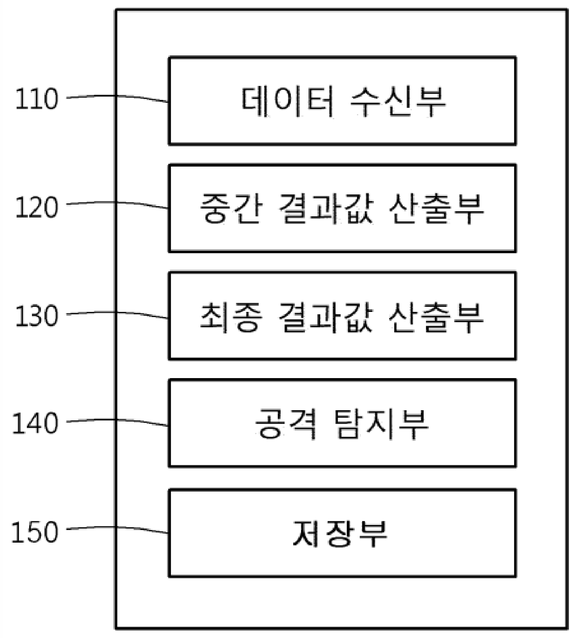
미해결 과제(Unmet needs)

- 기존 딥러닝 기반의 적대적 공격 탐지 기법의 한계
  - 딥러닝 모델 또는 기계학습 모델에 예기치 않은 데이터를 입력하여 시스템이 의도하지 않는 결과를 출력하도록 하는 적대적 공격에 대한 기존 탐지 기법의 경우, 모델을 최초 학습할 때 기본 모델과 공격 탐지 모델 각각에 대하여 학습(즉, 두 번의 학습)하여야 하고, 내부 구조를 달리하거나 알고리즘을 달리하는 등의 노력이 필요함
  - 또한, 모델을 업데이트할 경우에 또 다시 두 개의 모델을 별도로 업데이트 하여야 하며, 공격 탐지 모델을 보관하는 공간이 필요하여, 하나의 모델에서 기본 모델이 제공하는 기능과 공격 탐지 기능을 모두 제공할 수 있는 방법이 필요한 실정임

기술적 해결수단(발명의 구성)

- 1) 본 기술에 따른 적대적 공격 탐지 장치의 구성
  - 본 기술의 적대적 공격 탐지 장치(10)는 데이터 수신부(110), 중간 결과값 산출부(120), 최종 결과값 산출부(130) 및 공격 탐지부(140)로 구성되며, 저장부(150)를 포함할 수도 있음
  - 데이터 수신부는 유무선 통신망을 통하여 입력 데이터를 수신하거나, 소정의 입력 인터페이스를 통해 사용자로부터 입력 데이터를 수신함
  - 중간 결과값 산출부는 입력 데이터에 대한 탐지 모델의 적어도 하나의 중간 결과값을 산출하며, 최종 결과값 산출부는 입력 데이터에 대한 탐지 모델의 최종 결과값을 산출함
  - 공격 탐지부는 적어도 하나의 중간 결과값과 최종 결과값의 비교를 통해 적대적 공격 발생 여부를 판단함
  - 저장부에는 입력 데이터, 탐지 모델(기본 모델을 포함할 수 있음), 적어도 하나의 중간 결과값, 최종 결과값, 공격 탐지의 결과, 공격 탐지 과정에서 일시적으로 또는 비일시적으로 생성되는 데이터 등이 저장됨

본 기술에 따른 적대적 공격 탐지 장치의 구성

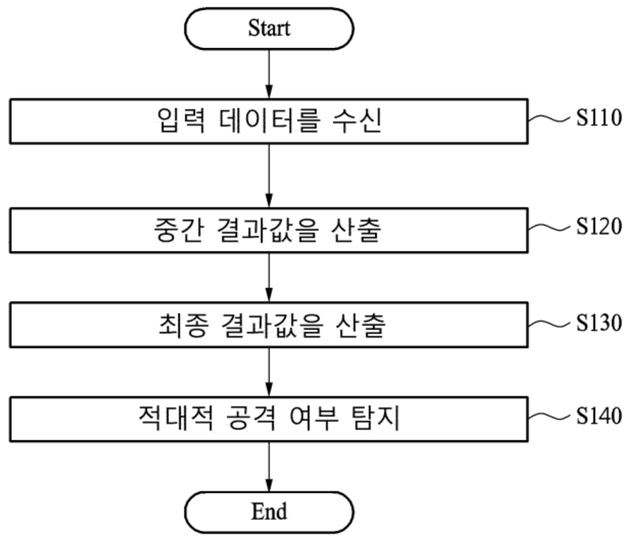


본 기술의 우수성 및 파급 효과

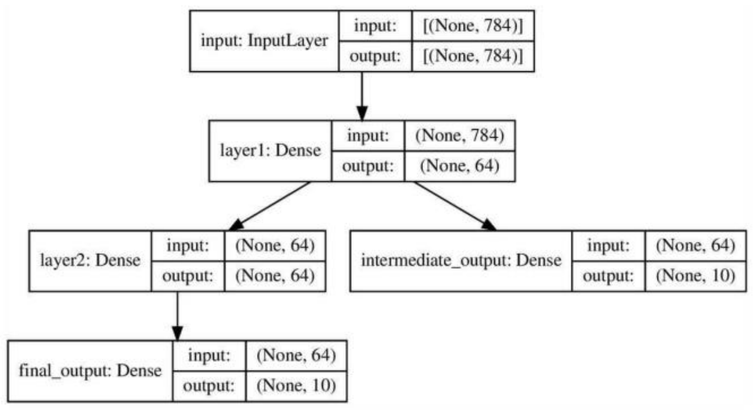
본 기술의 우수성(효과)

- 딥러닝 기반의 적대적 공격 탐지 모델
  - 본 기술의 적대적 공격 탐지 장치의 탐지 모델은 사전 학습되어 미리 정해진 동작을 수행하는 기본 모델과 중간 결과값을 출력하는 적어도 하나의 중간 출력 레이어로 구성됨
  - 기본 모델은 복수의 레이어들(입력 레이어, 복수의 중간 레이어, 및 출력 레이어)를 포함하는 딥러닝 모델 또는 기계 학습 모델을 말함
  - 입력 레이어는 입력 데이터를 복수의 중간 레이어들 중 첫번째 중간 레이어에 전송하는 역할을 수행하며, 출력 레이어는 복수의 중간 레이어들 중 마지막 중간 레이어의 출력을 입력받아 최종 결과값을 출력함
  - 중간 출력 레이어는 기본 모델에 포함된 복수의 중간 레이어들 중 어느 하나의 중간 레이어의 출력을 수신하여, 이에 기초한 중간 결과값을 출력함
  - 출력 레이어의 출력값(최종 결과값)과 중간 출력 레이어의 출력값(중간 결과값)이 동일한 경우 적대적 공격이 발생하지 않은 것으로 판단하고, 최종 결과값과 중간 결과값이 동일하지 않은 경우 적대적 공격이 발생한 것으로 판단함
  - 따라서, 기본 딥러닝 모델이 탐지 기능을 수행하여 모델 구조를 달리하거나 알고리즘을 달리할 필요가 없으며, 탐지 모델을 보관할 별도의 공간을 마련할 필요가 없음

본 기술에 따른 적대적 공격 탐지 방법



본 기술에 따른 적대적 공격 탐지 모델



적용 제품 및 파급 효과

- 보안 장치
- 본 기술을 통해 내부 구조나 알고리즘을 달리한 공격 탐지 모델을 별도로 구비하여 적대적 공격을 방어하는 기존 탐지 기법의 문제점을 해소할 수 있음

지식재산권 현황

발명의 명칭	출원/등록번호	출원/등록일자
딥러닝 기반의 적대적 공격 탐지 장치 및 방법	10-2609945	2023.11.30.
패밀리 특허 현황	**패밀리 국가**	
-	-	

문 의 | 고려대학교 산학협력단 기술사업화센터 권성진 02-3290-5832, giannie77@korea.ac.kr