정보보호론 #10

IoT 보안 및 AI보안

Prof. Byung Il Kwak

Review

- 악성코드
- 바이러스
- 웜
- 트로이 목마
- PUP
- 악성코드 탐지 및 대응책

Contents

- IoT 보안
- AI 에 대한 이해
- AI의 취약점 유형과 대안
- AI를 이용한 보안

CONTENTS

- IoT 보안

IoT 보안

- IoT의 개념
  - 인터넷에 연결되는 것을 의미
  - MIT Auto-ID 센터 설립자인 케빈 애시턴이 1999년 사물 인터넷의 개념과 용어를 처음으로 제안
  - 이미 상용화 된 IoT 기기로는 전구, 자동차, 냉장고, 보일러, 자물쇠, 칫솔, 개인 비서, 프린터 등
  - 매우 다양하지만, 물건에 시스템을 결합한 형태라는 점은 동일

IoT 보안

- 일상생활에서 사용되는 IoT 제품 유형별 주요 보안 위협

유형	주요 제품	주요 보안 위협	주요 보안 위협 원인
멀티미디어 제품	스마트 TV, 스마트 냉장고 등	• PC 환경의 모든 악용 행위 • 카메라/마이크 내장 시 사생활 침해	• 인증 매커니즘 부재 • 강도가 약한 비밀번호 • 펌웨어 업데이트 취약점 • 물리적 보안 취약점
생활 가전 제품	청소기, 인공지능 로봇 등	• 알려진 운영체제 취약점 및 인터넷 기반 해킹 위협 • 로봇 청소기에 내장된 카메라를 통해 사용자 집 모니터링	• 인증 매커니즘 부재 • 펌웨어 업데이트 취약점 • 물리적 보안 취약점
네트워크 제품	홈캠, 네트워크 카메라 등	• 사진 및 동영상을 공격자의 서버 및 이메일로 전송 • 네트워크에 연결된 홈캠 등을 원격으로 제어하여 임의 촬영하는 등 사생활 침해	• 접근 통제 부재 • 전송 데이터 보호 부재 • 물리적 보안 취약점
제어 제품	디지털 도어락, 가스 밸브 등	• 제어 기능 탈취로 도어락 임의 개폐	• 인증 매커니즘 부재 • 강도가 약한 비밀번호 • 접근 통제 부재 • 물리적 보안 취약점
	모바일 앱(웹) 등	• 앱 소스 코드 노출로 IoT 제품 제어 기능 탈취	• 인증 정보 평문 저장 • 전송 데이터 보호 부재
센서 제품	온 · 습도 센서 등	• 잘못되거나 변조된 온 · 습도 정보 전송	• 전송 데이터 보호 부재 • 데이터 무결성 부재 • 물리적 보안 취약점

출처: 한국인터넷진흥원(KISA)

IoT 보안

- IoT의 개념
  - IoT 공통 보안 7대 원칙
    - Security by Design
    - Privacy by Design
    - 안전한 소프트웨어와 하드웨어 개발 기술 적용 및 검증
      - 시큐어 코딩
      - 소프트웨어 보안성 검증
      - 시큐어 하드웨어 장치 활용
      - 소프트웨어 보안 기술과 하드웨어 보안 기술 융합
    - 안전한 초기 보안 설정 방안 제공
    - 보안 프로토콜 준수와 안전한 파라미터 설정
    - IoT 제품과 서비스의 취약점 보안 패치 및 업데이트 지속 이행
    - 안전한 운영을 위한 정보 보호 및 프라이버시 관리 체계 마련
    - IoT 침해 사고 대응 체계 및 책임 추적성 확보 방안 마련

CONTENTS

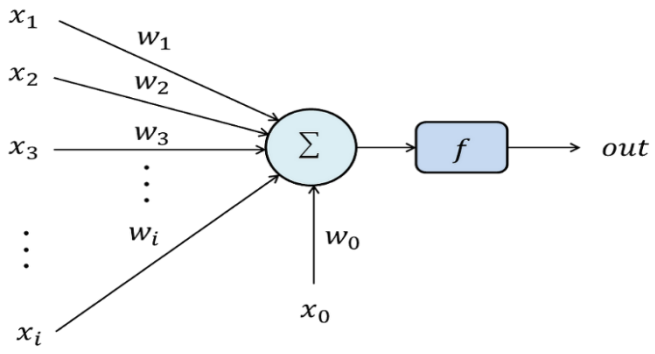
- AI 에 대한 이해

AI 에 대한 이해

- AI의 역사
  - AI의 시작
    - 1943년 논리학자인 월터 피츠와 신경외과의인 워렌 맥컬럭의 논문에서 최초의 인간 두뇌에 관한 모델이 등장
    - 1950년 영국의 수학자 앨런 튜링의 논문에서 '생각하는 기계'에 대해 기술
    - 1956년 존 매카시 교수가 '다트머스 AI 컨퍼런스'를 개최하면서 초청장 문구에 'AI'라는 용어를 처음으로 사용

AI 에 대한 이해

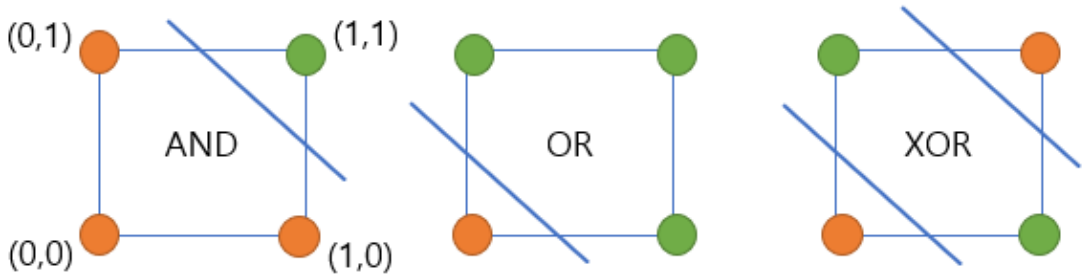
- AI의 역사
  - AI의 발전
    - 1950년대의 인공지능 연구는 크게 기호주의와 연결주의의 두 분야로 전개
      - 기호주의: 인간의 지능과 지식을 기호화, 매뉴얼화하는 접근법
      - 연결주의: 1943년 월터 피츠와 워렌 맥컬럭이 연구한 뇌 신경 네트워크의 재현을 목표로 하는 접근법
    - 퍼셉트론은 인간의 사진을 대상으로 남자와 여자를 구별해 내면서 뉴욕 타임즈에 등장
      - 퍼셉트론: 인공 신경망(딥 러닝)의 기본이 되는 알고리즘



[Source https://untitledtblog.tistory.com/27]

AI 에 대한 이해

- AI의 역사
  - AI의 빙하기
    - 단층 퍼셉트론의 한계가 수학적으로 증명됨
    - 퍼셉트론이 무너지고 설상가상으로 2년 뒤인 1971년에 로젠블랫이 사망. 신경망 열기가 급격히 냉각
    - 퍼셉트론으로는 XOR 같은 비선형 문제는 해결할 수 없음



AI 에 대한 이해

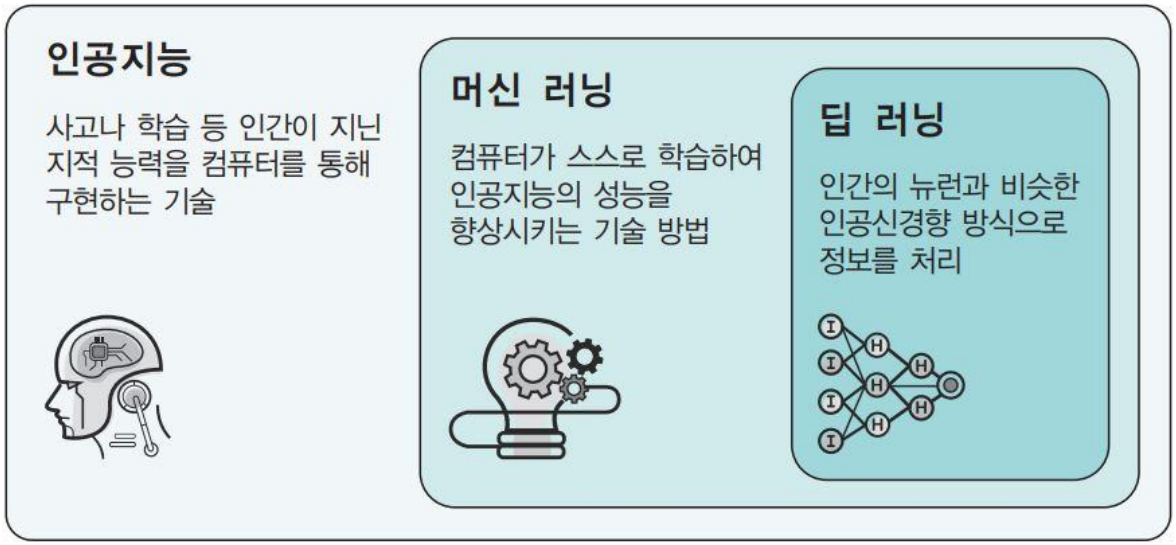
- AI의 역사
  - AI의 부활
    - 제프리 힌튼 교수는 다층 퍼셉트론, MLP과 역전파 알고리즘을 실험을 통해 증명하여 XOR 문제를 해결
    - 하버드대학 폴 워보스가 다층 퍼셉트론 환경에서 학습을 가능하게 하는 역전파 알고리즘으로 박사 학위 논문 발표
    - 1986년 데이빗 럼멜하트와 제프리 힌튼이 최적의 신경망 변수들을 찾아내는 적합을 증명
  - AI의 2차 빙하기
    - 기울기 소실과 과적합문제로 2차 빙하기를 맞이함
    - 기울기 소실
      - 다층 신경망의 은닉층을 늘리면 신경망의 깊이가 깊어질수록 오히려 기울기가 사라져 학습이 되지 않는 문제
    - 과적합
      - 신경망이 깊어질수록 너무 정교한 패턴을 감지하게 되어 새로운 데이터에 대해서는 정확성이 떨어지는 문제

AI 에 대한 이해

- AI의 역사
  - 딥 러닝의 시작
    - 제프리 힌튼은 가중치의 초기값을 제대로 설정한다면 깊은 신경망을 통한 학습이 가능하다는 것을 밝혀냄
    - 인공 신경망이라는 단어가 들어간 논문을 학회에 투고하면 제목만 보고 거절당하거나 사람들의 관심을 끌지 못하기 때문에 제프리 힌튼은 이 논문에 deep을 붙인 DNN (Deep Neural Network)이라는 용어를 사용

AI 에 대한 이해

- AI 기술의 분류
  - AI가 곧 딥 러닝은 아니며 아직 AI 관련 용어에 대한 정의가 명확하지 않고 혼용되는 경우가 많음

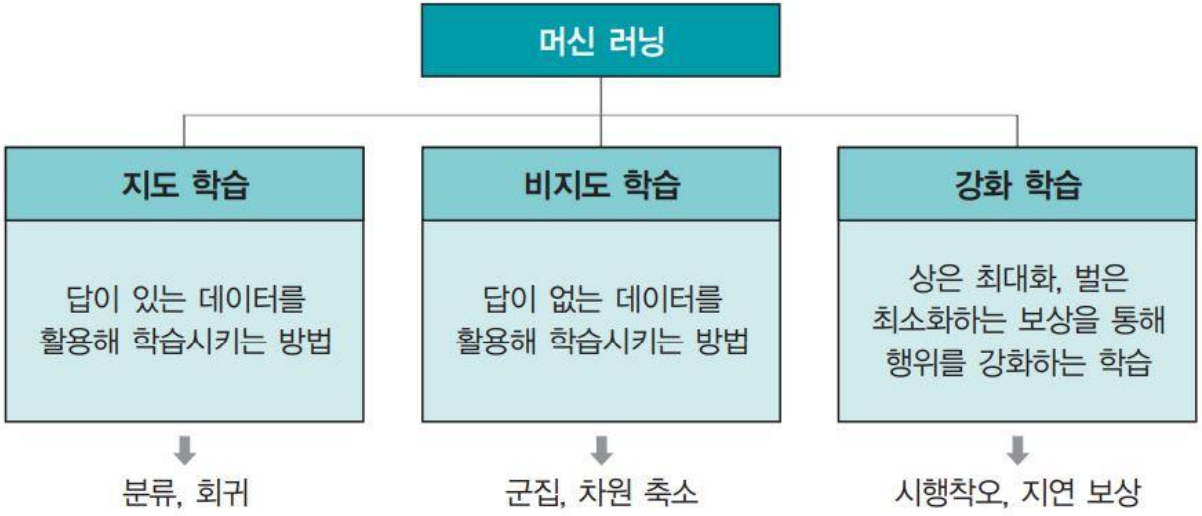


인공지능은 학습, 문제 해결, 패턴 인식 등과 같이 주로 인간 지능과 연결된 인지 문제를 해결하는데 주력하는 컴퓨터 공학 분야

- AWS 에서의 AI 정의 -

AI 에 대한 이해

- AI 기술의 분류
  - 머신 러닝의 분류
    - 머신 러닝 역시 상당 부분 통계학적인 기술을 포함
      - 지도 학습: 분류나 회귀에 사용
      - 비지도 학습: 군집에 사용
      - 강화 학습: 환경에서 취하는 행동에 대한 보상을 이용하여 학습을 진행



AI 에 대한 이해

- AI 기술의 분류
  - 지도 학습 (Supervised Learning)
    - 답이 있는 데이터를 활용해 학습시키는 방법
    - 입력 값(X) 이 주어지면 입력 값에 대한 라벨(Y)을 주어 학습
    - 분류 (Classification)
      - 주어진 데이터를 정해진 레이블(범주)에 따라 나누는 것
      - 범주가 2개 이면 이진 분류 , 3개 이상이면 다중 클래스 분류
    - 회귀 (Regression)
      - 어떤 데이터들의 특징을 토대로 값을 예측하는 것
      - 결과 값은 실수 값을 가질 수 있음

AI 에 대한 이해

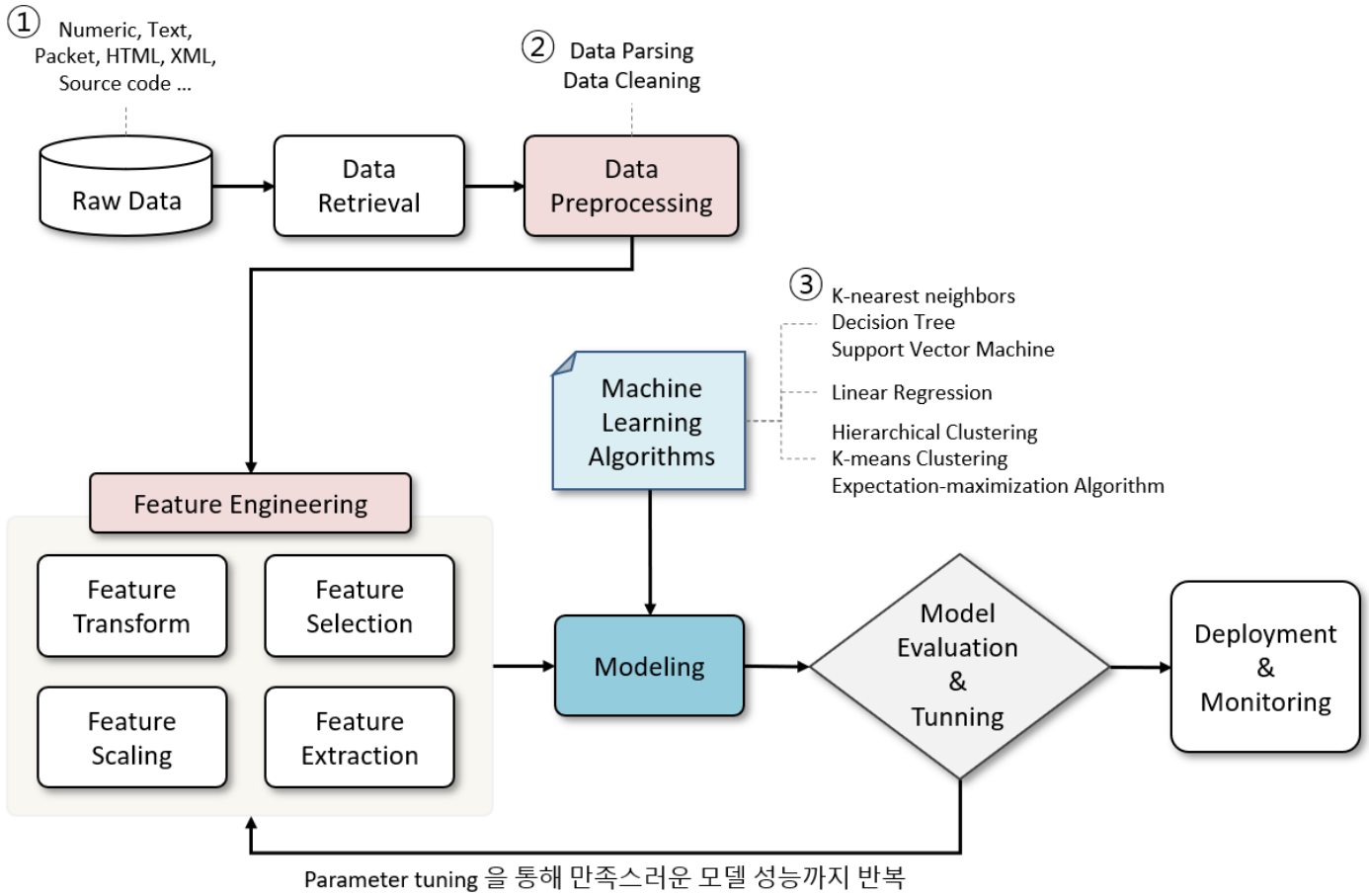
- AI 기술의 분류
  - 비지도 학습
    - 지도 학습과 달리 답이 없는 데이터를 비슷한 특징끼리 군집화하여 새로운 데이터에 대한 결과를 예측하는 것
    - 군집
      - 특정 기준에 따라 유사한 특성의 데이터를 각각의 그룹으로 분류
      - 고유한 패턴 또는 특성을 찾기 위해 클러스터링을 사용
    - 차원 축소
      - 많은 변수들 중 유의미한 변수들을 식별하여 개수를 줄이는 작업
      - 일부 변수가 중복되거나 작업과 아무 관련이 없는 경우가 많기 때문
      - 변수를 줄이면 잠재되어 있는 진정한 관계를 도출하기가 용이

AI 에 대한 이해

- AI 기술의 분류
  - 강화 학습
    - 지도·비지도 학습과는 다른 종류의 알고리즘
    - 학습하는 시스템을 '**에이전트**'라고 부르며, 환경을 관찰해서 행위를 수행하고 보상을 받음
    - 시간이 지나면서 가장 큰 보상을 얻기 위해 '정책'이라고 부르는 최상의 전략을 스스로 학습
    - 정책은 주어진 상황에서 에이전트가 어떻게 행동해야 하는지를 판단하는 것
    - 기계는 최대의 보상을 산출하는 행위를 발견하기 위해 서로 다른 시나리오를 시도
    - 시행착오와 지연 보상은 다른 기법과 구별되는 강화 학습만의 특징

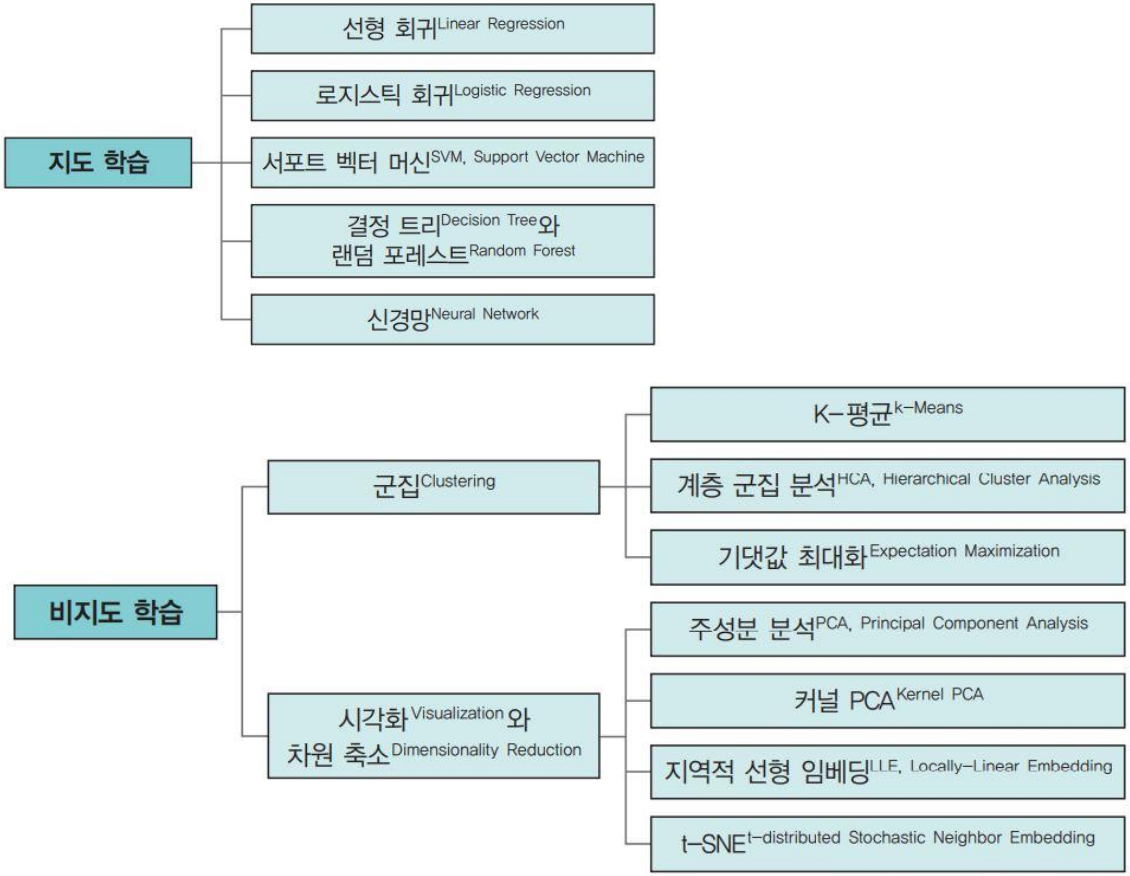
AI 에 대한 이해

- AI 기술의 분류
  - 머신러닝



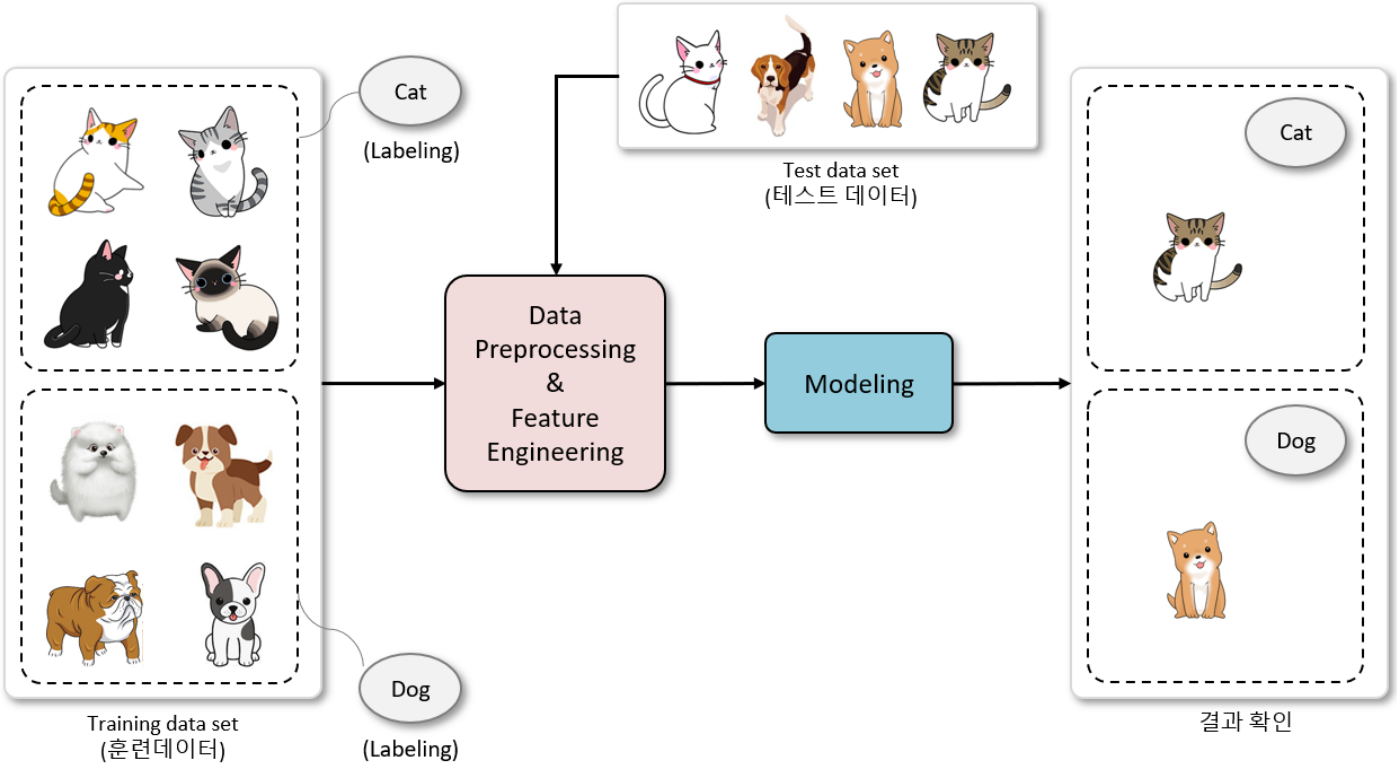
AI 에 대한 이해

- AI 기술의 분류
  - 지도학습 / 비지도학습



AI 에 대한 이해

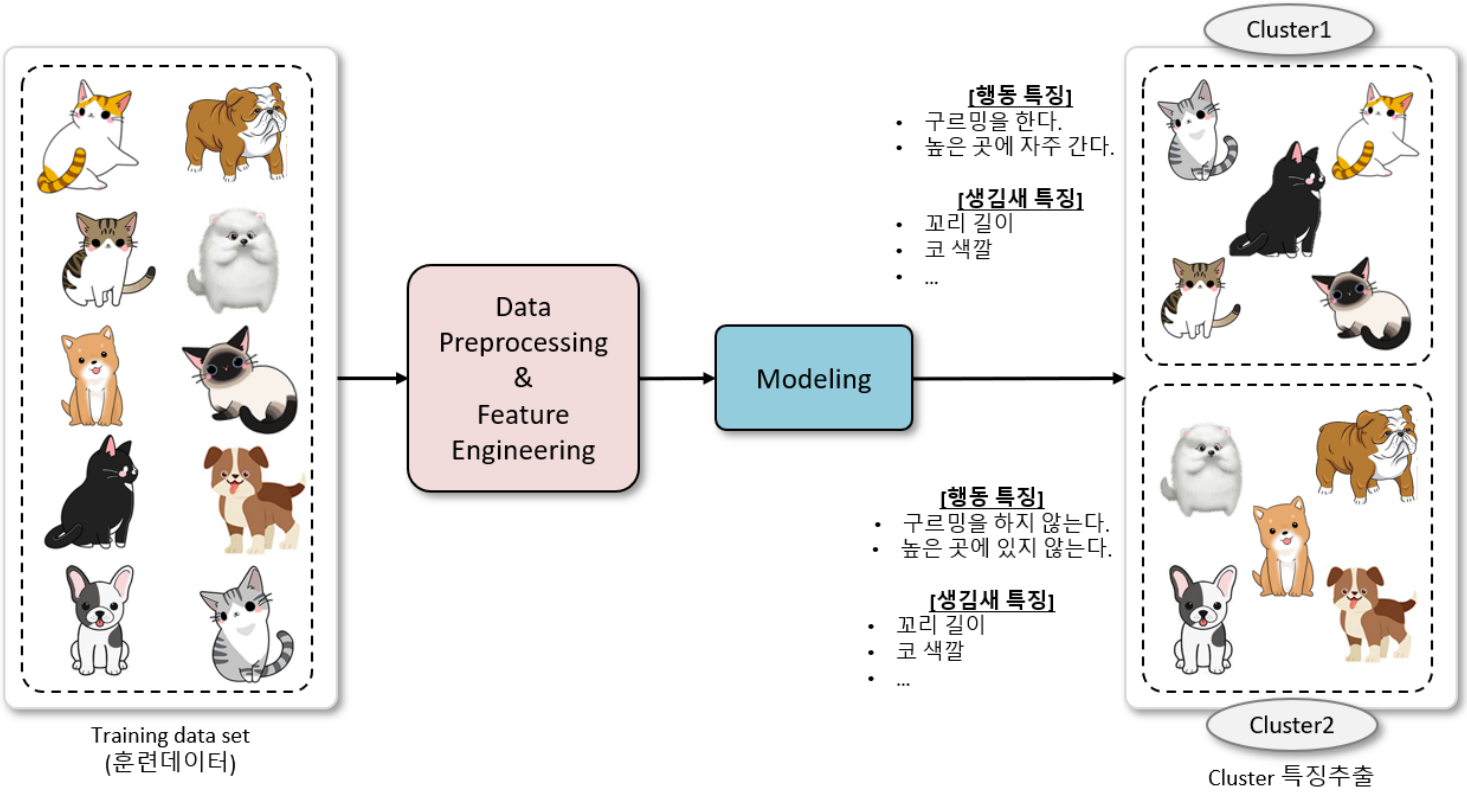
- AI 기술의 분류
  - 지도학습 - 정답지 (Labeling)이 있는 데이터를 대상으로 학습하는 과정



AI 에 대한 이해

AI 기술의 분류

- 정답지 (Labeling)이 없는 데이터를 비슷한 특징끼리 군집화하여 새로운 결과를 예측하는 학습과정
- 비지도학습은 답을 맞히는 목적으로 학습하지는 않음

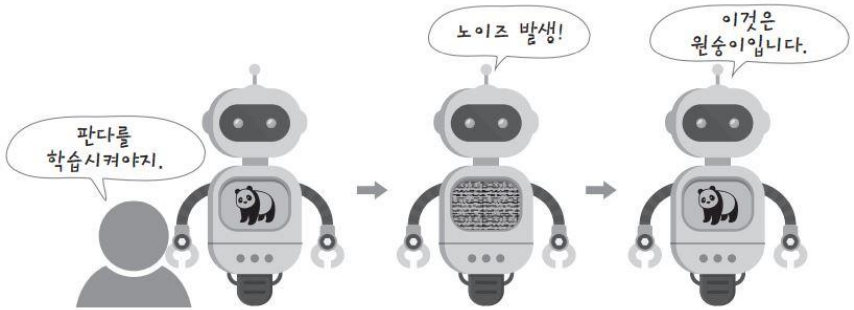


CONTENTS

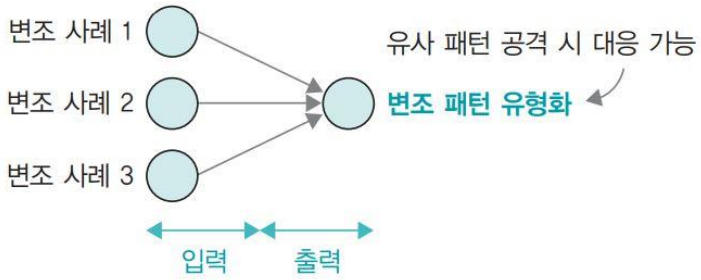
- AI의 취약점 유형과 대안

AI의 취약점 유형과 대안

- 데이터 변조 공격과 대안
  - 회피 공격
    - 학습 과정에서 데이터에 무작위의 오류가 존재하는 노이즈를 고의적으로 추가하면 인공지능은 다른 이미지로 판단
    - 기존 해킹 방법이 유무선 네트워크나 단말기의 취약점을 이용, 인공지능에 대한 해킹은 인공지능 자체의 취약점을 이용
    - **(대응방안) 데이터가 변조되었다면 변조 공격을 학습 데이터에 포함해 훈련시키는 방법으로 대응할 수 있음**



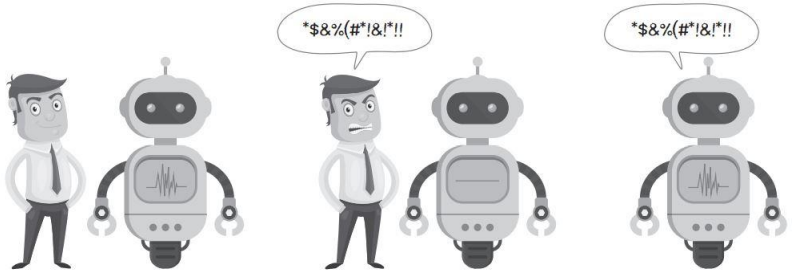
인공지능 판단 오류



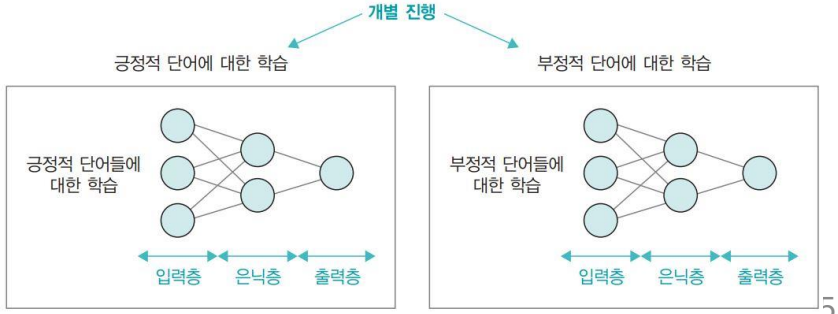
데이터 변조 사례에 대한 데이터 학습

AI의 취약점 유형과 대안

- 데이터 변조 공격과 대안
  - 악의적 데이터 주입 공격
    - 중독 공격
      - 악의적인 데이터를 이용해 인공지능 시스템이 오작동을 일으키도록 하는 공격
        - 예) 스캐터랩의 '**이루다**'는 일부 사용자들이 이루다에게 **악의적인 발언들을** (욕설, 인종 차별 및 성 차별 발언, 정치적 발언 등) **학습**시키면서 정상적인 서비스가 불가능해짐
      - **(대응방안) 부정적인 데이터에 대한 사전 학습으로 대응 가능**
      - **또한, 사전 학습된 단어들과 비교하여 추가 학습이 불가하거나 답변을 우회하도록 프로그램을 설계하는 방법도 고려**



악의적 단어를 학습하여 부정적 단어를 생성하는 인공지능



긍정적인 단어와 부정적인 단어에 대한 개별 학습

AI의 취약점 유형과 대안

- 데이터 추출 공격
  - 전도 공격
    - 데이터 추출 공격은 인공지능에서 사용하는 데이터 자체를 탈취하는 공격
    - 인공지능에 수많은 쿼리를 한 후, 산출된 결과를 분석해 인공지능에서 사용된 데이터를 추출하는 공격
    - 데이터 추출 공격에는 질의 횟수를 조정하는 것으로 대응할 수 있음(완벽한 대응방안은 될 수 없음)

사용자당 질의 횟수 제한

CONTENTS

- AI를 이용한 보안

AI의 취약점 유형과 대안

- 스팸 메일 탐지
  - 나이브 베이즈 분류기
    - 최초의 스팸 메일 솔루션에 일반적으로 적용된 알고리즘
  - 베이즈의 정리
    - P(A)가 A가 일어날 확률, P(B)가 B가 일어날 확률
    - P(B|A)는 A가 일어난 뒤 B가 일어날 확률
    - P(A|B)는 B가 일어난 뒤 A가 일어날 확률

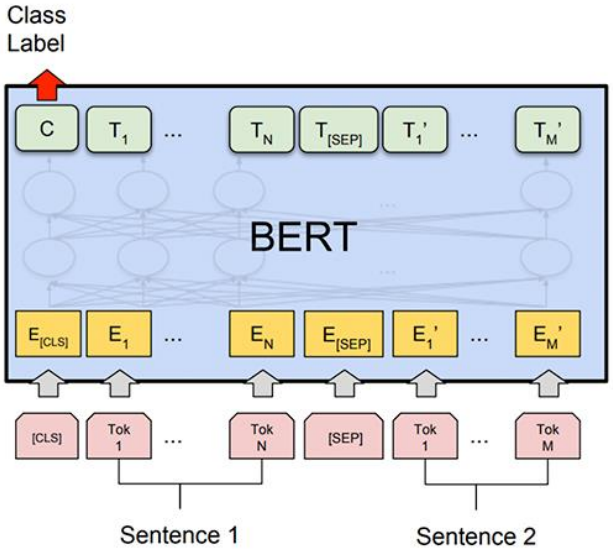
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

    - 입력 텍스트가 정상 / 스팸 메일 구분을 위한 확률
      - P(정상 메일|텍스트) = (P(텍스트|정상 메일) * P(정상 메일)) / P(텍스트)
      - P(스팸 메일|텍스트) = (P(텍스트|스팸 메일) * P(스팸 메일)) / P(텍스트)
    - P(정상 메일 | 텍스트)가 P(스팸 메일 | 텍스트)보다 크다면 정상 메일이라고 볼 수 있으며, 그 반대라면 스팸 메일

AI의 취약점 유형과 대안

- 스팸 메일 탐지
  - BERT
    - 구글에서 개발한 NLP 사전 훈련 기술로 모든 자연어 처리 분야에서 좋은 성능을 내는 범용 언어 모델
    - 딥러닝에서 인공 신경망의 뉴런과 유사 역할을 하는 어텐션 알고리즘을 적용함
    - 방대한 양의 텍스트를 이용하여 모델을 훈련
    - 최근 BERT를 활용한 스팸 메일 솔루션은 높은 정확도를 보임

Google
BERT



AI의 취약점 유형과 대안

- 네트워크 침입 탐지
  - 전문가 시스템
    - 특정 응용 분야 전문가의 지식 및 능력을 체계적으로 잘 조직하여 컴퓨터 시스템에 입력해 해당 분야의 비전문가라도 전문가에 상응하는 능력을 발휘할 수 있도록 쉽고 빠르게 도움을 주는 시스템
    - 전문가 시스템은 앞서 AI의 역사에서 언급한, 논리적인 체계로 문제를 푸는 기호주의의 한 분야에 속함
  - 오탐률
    - 거짓 양성 **'Type 1 에러'** 라고 하며, **실제로는 공격이 아닌데 공격이라고 탐지하는 것**
    - 기존의 전문가 시스템은 이와 반대 로 거짓 음성 **'Type 2 에러'** 가 상대적으로 높음
      - Type 2 에러는 **공격을 받았으나 이를 탐지하지 못하는 것**

AI의 취약점 유형과 대안

- 악성 코드 탐지
  - 기존에는 악성 코드를 탐지하기 위해 악성 코드의 일부분을 매칭해보거나 특정 부분의 해시 값을 생성하여 비교해보는 등의 방법이 사용
  - 최근에는 이런 탐지 방법을 회피한 고도화된 방식으로 다양한 신종 및 변종 악성 코드가 나타나 탐지가 어려움

AI의 취약점 유형과 대안

- 악성 코드 탐지
  - 현재는 프로그램이 지닌 일반적인 특징들을 변수화하여 이를 기반으로 악성 코드와 정상 코드를 머신러닝 모델에 학습시켜 악성 코드를 탐지하는 방법이 제안 및 연구
  - 최근에 응용 프로그램의 행위에 기반한 특징을 분석해서 이를 변수로 두고 악성 코드를 탐지하는 기술이 적용되고 있는데, 이는 상당한 정확도를 보여줌
  - 변수로 활용되는 데이터의 특징

특징	설명
API Call	API는 응용 프로그램의 수행 작업을 파악할 수 있으며, 간접적으로 행위의 의도를 추론할 수 있다.
Runtime Log	입출 네트워크 데이터, 파일 읽기 및 쓰기 작업, DexClassLoader를 통한 서비스 시작, 로드된 클래스, 네트워크를 통한 정보 유출, 파일 및 SMS 정보를 포함한다.
시스템 자원	네트워크 액세스, 연락처, SMS 송수신 기능 같은 정보를 포함한다. 응용 프로그램이 리소스를 사용할 경우 개발자는 컴파일할 때 해당 리소스와 관련된 권한을 요청한다.
네트워크	악성 코드가 발생시킨 패킷의 시작과 종료 시간, 송 · 수신 패킷의 플로, IP 주소, Port 번호 등의 정보를 통해 악성 코드를 탐지한다.

AI의 취약점 유형과 대안

- CCTV
  - 물리적 보안에서 중요한 역할을 담당하는 CCTV에도 머신 러닝이 적용
  - CCTV에 찍히는 영상을 AI 기술로 실시간 처리하여 무단 침입과 같은 침해 사고를 감지
  - 다국적 정유 기업인 쉘은 각 주유소에 설치된 CCTV로 모니터링하며 이 영상 데이터를 Azure클라우드 환경에서 분석하여 주유소와 관련한 위험 요인을 탐지

Summary

- IoT 보안
- AI 에 대한 이해
- AI의 취약점 유형과 대안
- AI를 이용한 보안

참고문헌

- 정보 보안 개론 - 한권으로 배우는 핵심 보안 이론, 양대일, 한빛 아카데미

Q & A