11주차: IoT 보안과 AI 보안

ChulSoo Park

School of Computer Engineering & Information Technology

Korea National University of Transportation

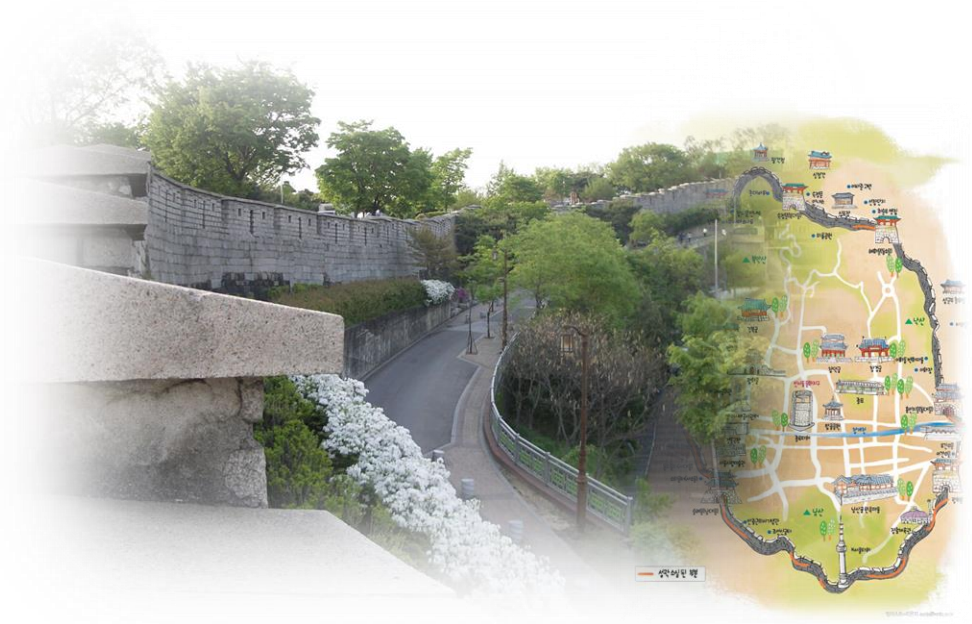
E-Mail : pcs8321@naver.com

학습목표 (11주차)

- IoT 보안의 이해
- AI 보안의 이해
- AI의 역사와 종류별 원리 이해
- AI의 취약점 파악
- AI의 보안 적용 사례 학습

10 CHAPTER

IoT 보안과 AI 보안



CONTENTS

1. IoT 보안
2. AI의 이해
3. AI의 취약점 유형과 대안
4. AI를 이용한 보안

1. IoT 보안

자율주행 택시 상용화(서울 상암동) 사례

1. IoT 보안

커넥티드 카 콘텐츠 서비스 사례

이름

- LM1903170201_19v1_커넥티드카 콘텐츠 서비스 기획(편집완료)_1104.hwp
- LM1903170202_19v1_커넥티드카 콘텐츠 서비스 UI UX 디자인 개발(편집완료)_1106.hwp
- LM1903170203_19v1_커넥티드카 콘텐츠·서비스 운영 플랫폼 구성(편집완료)_1104.hwp
- LM1903170204_19v1_커넥티드카 차량 서비스 운용(편집완료)_1106.hwp
- LM1903170205_19v1_커넥티드카_차량데이터처리기술개발(편집완료)_1110.hwp
- LM1903170206_19v1_커넥티드카 콘텐츠 서비스 보안기술 개발(편집완료)_1106.hwp
- LM1903170207_19v1_커넥티드카 콘텐츠 서비스 입력기술 구현(편집완료)_1106.hwp
- LM1903170208_19v1_커넥티드카 콘텐츠 서비스 표출기술 구현(편집완료)_1107.hwp

NCS학습모듈 개발이력			
발행일	2021년 12월 31일		
세분류명	커넥티드카콘텐츠서비스(19031702)		
개발기관	한국정보통신기술사협회(개발책임자: 박철수), 한국직업능력연구원		
집필진	김영욱(현대자동차)*	감수진	강상미(현대HDS)
	권오신(금융결제원)		김준삼(매경닷컴)
	권영란((주)레몬헬스케어)		윤주희(프리랜서)
	박은경(랩스마트)		현지은(동일부)
	박재성(DGB생명)		
	박수용((주)아이티아이즈)		
	이상아((주)아이리포)		
	하광림((주)씨에스리)		*표시는 대표집필자임

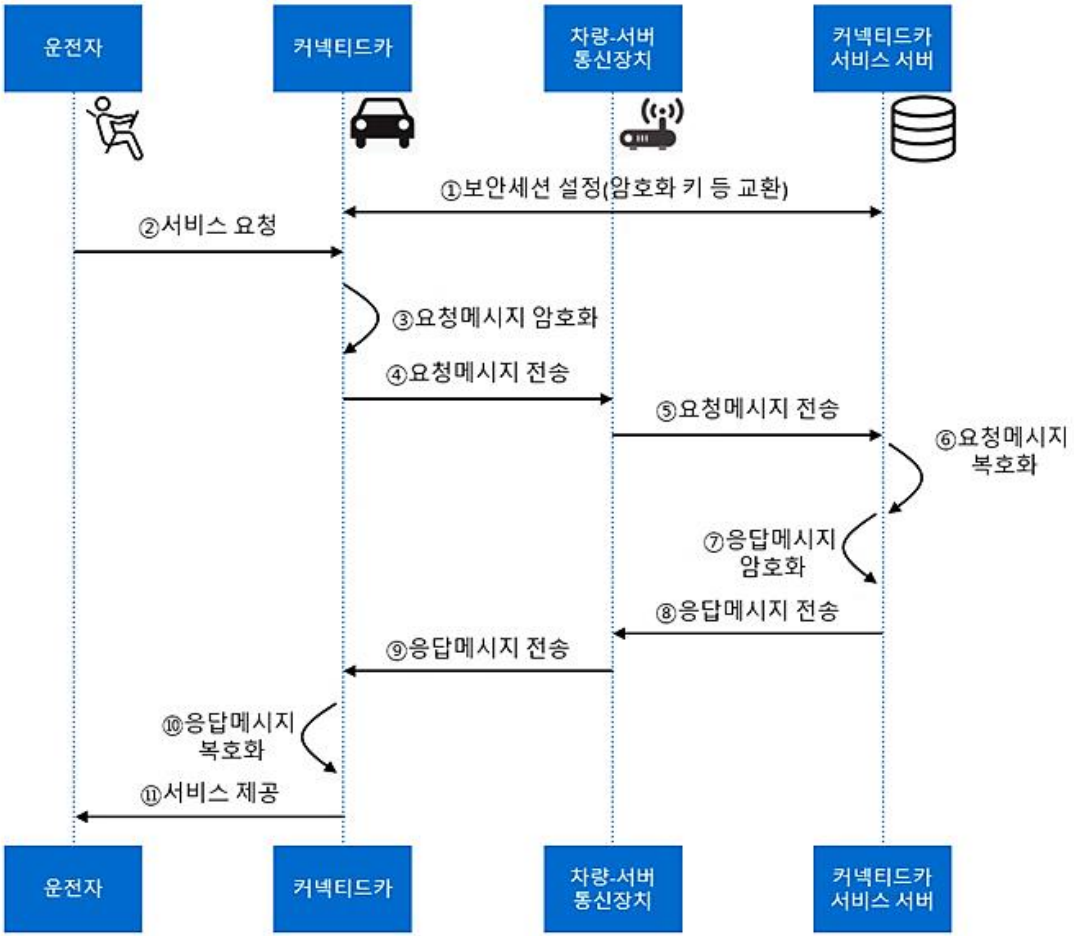
커넥티드카 콘텐츠·서비스 보안기술 개발(LM1903170206_19v1)	
저작권자	교육부
연구기관	한국직업능력연구원
발행일	2021. 12. 31.

※ 이 학습모듈은 자격기본법 시행령(제8조 국가직무능력표준의 활용)에 의거하여 개발하였으며, NCS통합포털사이트(http://www.ncs.go.kr)에서 다운로드 할 수 있습니다.

차 례

1. IoT 보안

커넥티드 카 콘텐츠 서비스 사례



출처: 집필진 제작(2021)
[그림 4-6] 커넥티드카의 종단 간 암호화 보안 요소기술을 개발하기 위한 흐름도

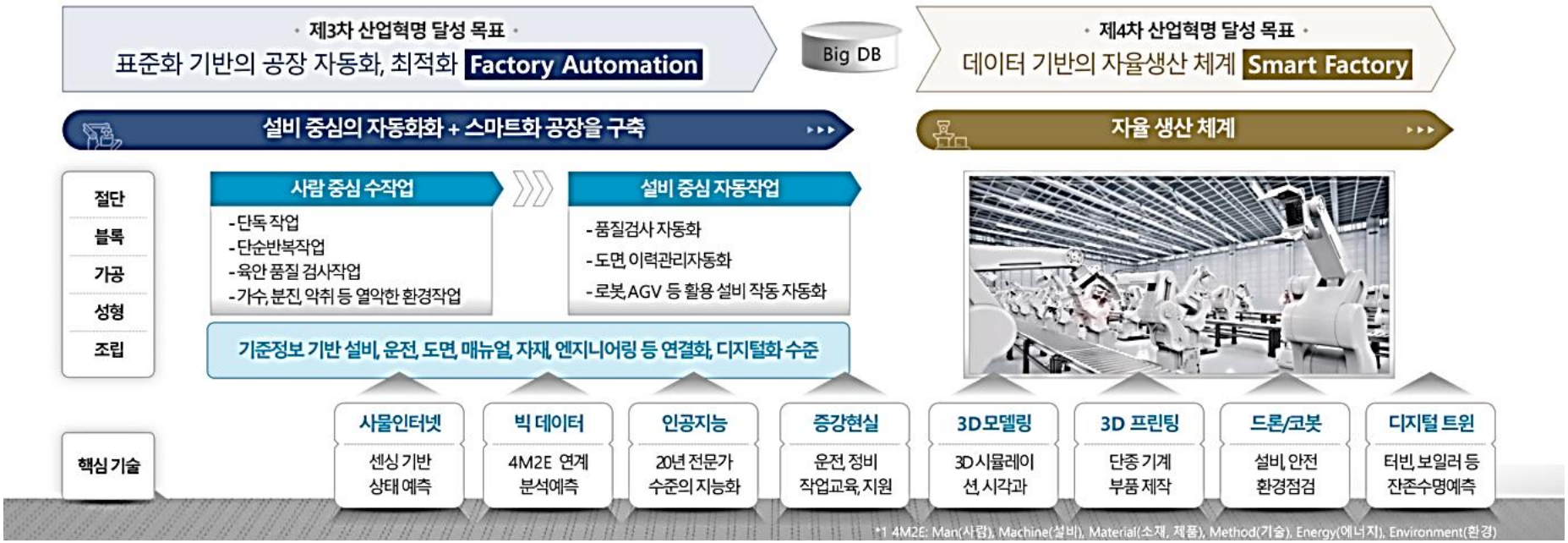
1. IoT 보안

스마트 팩토리 IoT 사례 / 공장 랜섬웨어로 stop 사례

3. 추진전략 – 노동 집약산업을 자동화후 스마트 공장으로

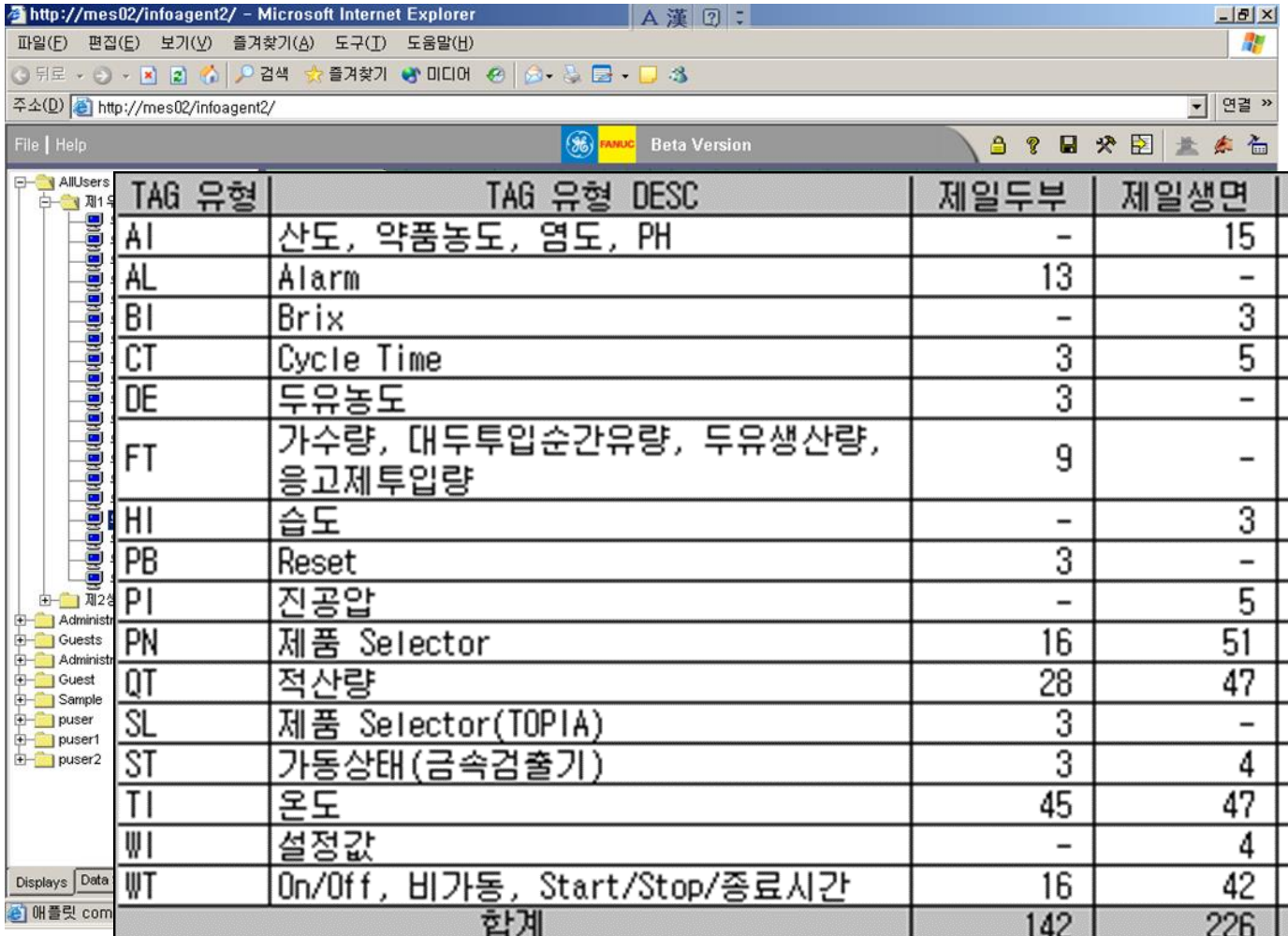
KOSMO KOREA SMART MANUFACTURING OFFICE

◆ 3차 산업혁명의 과업(자동화)을 완성할 수 있도록 기준정보체계, 4M2E*1 관리체계를 분석하여 자동화 + 무인화를 달성하고 빅 데이터베이스를 구축하면서 스스로 고객이 원하는 제품을 생산하는 자율생산 체계로 글로벌 시장 장악



1. IoT 보안

풀무원 사례 : 박광순 전무님



TAG 유형	TAG 유형 DESC	제일두부	제일생면	명가식품	제이생면
AI	산도, 약품농도, 염도, PH	-	15	-	-
AL	Alarm	13	-	-	-
BI	Brix	-	3	-	-
CT	Cycle Time	3	5	-	-
DE	두유농도	3	-	-	-
FT	가수량, 대두투입순간유량, 두유생산량, 응고제투입량	9	-	-	-
HI	습도	-	3	3	-
PB	Reset	3	-	-	-
PI	진공압	-	5	-	-
PN	제품 Selector	16	51	32	12
QT	적산량	28	47	38	28
SL	제품 Selector(TOPIA)	3	-	-	-
ST	가동상태(금속검출기)	3	4	8	8
TI	온도	45	47	18	4
WI	설정값	-	4	-	-
WT	On/Off, 비가동, Start/Stop/종료시간	16	42	30	25
합계		142	226	129	77

4. AI를 이용한 보안

스마트 팩토리 사례

1. IoT 보안

B사 사례

□ 스마트공장 구축 주요 내용 및 목표수준

주요내용	목표 수준	
	현 수준	목표수준
1. 떡 판매 빅데이터 분석을 통한 판매수요예측(매장 소매판매량) 2. IIoT 기술을 활용한 떡 성형 중량 자동제어 시스템 구축 3. BOM 프로그램 개발을 통한 Back flush 적용 자재 재고관리 시스템 구축 4. 경영자 의사결정을 지원하기 위한 생산지표관리 프로그램 개발 1> 생산수율관리 2> 원가관리 3> 생산량그래프 5. IIoT 기술 활용 설비 신호취득을 통한 모니터링 시스템 구축 1> 떡 증자 설비(5곳) : 압력 및 온도 2> 떡 성형 설비(3대) : 성형기 속도, 오일분사시간, 떡누름속도, 떡 중량 등 3> 떡 포장 설비(3대) : 전원, 가동, 생산수량 4> 설탕, 소금 계량기(1대) : 중량 6. 기구축 시스템 모니터링 보완 7. 떡 생산이력 데이터를 실시간 수집하여 통계분석(SPC), Cpk 차트 분석 8. 자동화 설비 도입 및 개조 1> 떡 성형 설비 개조(자동제어가 가능하도록 개조) 2> 자동 포장설비 개조(무인작업 및 자동제어가 가능하도록 개조) 3> 자동 중량 측정 설비 도입(자동제어가 가능하도록 도입) 4> 설탕, 소금 계량기 도입(떡 품질관리를 위해 도입)	기초1	중간1

1. IoT 보안

IoT의 개념

- 인터넷에 연결되는 것을 의미
  - ✓ 사물인터넷(IoT, Internet of Things)은 각종 사물에 센서와 통신 기능을 내장하여 인터넷에 연결하는 기술
  - ✓ 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

1. IoT 보안

IoT의 유형별 보안 위협

• 제품 유형별 주요 보안위협 •

유형	주요 제품	주요 보안위협	주요 보안위협 원인
멀티미디어 제품	스마트TV, 스마트 냉장고 등	• PC 환경에서의 모든 악용 행위 • 카메라/마이크 내장 시 사생활 침해	• 인증 메커니즘 부재 • 강도가 약한 비밀번호 • 펌웨어 업데이트 취약점 • 물리적 보안 취약점
생활가전 제품	청소기, 인공지능 로봇 등	• 알려진 운영체제 취약점 및 인터넷 기반 해킹 위협 • 로봇청소기에 내장된 카메라를 통해 사용자 집 모니터링	• 인증 메커니즘 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 촬영 등 사생활 침해	• 접근통제 부재 • 전송데이터 보호 부재 • 물리적 보안 취약점
제어제품	디지털 도어락, 가스밸브 등	• 제어기능 탈취로 도어락의 임의 개폐	• 인증 메커니즘 부재 • 강도가 약한 비밀번호 • 접근통제 부재 • 물리적 보안 취약점
	모바일 앱(웹) 등	• 앱 소스코드 노출로 IoT 제품 제어기능 탈취	• 인증정보 평문 저장 • 전송데이터 보호 부재
센서 제품	온/습도 센서 등	• 잘못된 또는 변조된 온 · 습도 정보 전송	• 전송데이터 보호 부재 • 데이터 무결성 부재 • 물리적 보안 취약점

출처 : KISA 홈,가전 IoT 보안 가이드 p.13

1. IoT 보안

IoT 공통 보안 7대 원칙

원칙	내 용
1	정보보호와 프라이버시 강화를 고려한 IoT 제품/서비스 설계 - "Security by Design" 및 "Privacy by Design" 기본 원칙 준수
2	안전한 SW, HW 개발 기술 적용 및 검증 - 시큐어 코딩, 소프트웨어, 애플리케이션 보안성 검증 및 시큐어 하드웨어 장치 활용
3	안전한 초기 보안 설정 방안 제공 - "Secure by Default" 기본 원칙 준수
4	보안 프로토콜 준수 및 안전한 파라미터 설정 - 통신 및 플랫폼에서 검증된 보안 프로토콜 사용(암호/인증/인가 기술
5	IoT 제품/서비스의 취약점 보안 패치 및 업데이트 지속 이행 - S/W와 H/W의 보안 취약점에 대해 모니터링하고 업데이트 지속 수행
6	안전한 운영/관리를 위한 정보보호 및 프라이버시 관리 체계 마련 - 사용자 정보 취득-사용-폐기의 전주기 정보의 보호 및 프라이버시 관리
7	IoT 침해 사고 대응체계 및 책임 추적성 확보 방안 마련 - 보안 사고에 대비한 침입탐지와 사고 시 분석 및 책임 추적성 화보

출처 : 미래창조과학부,KISA

1. IoT 보안

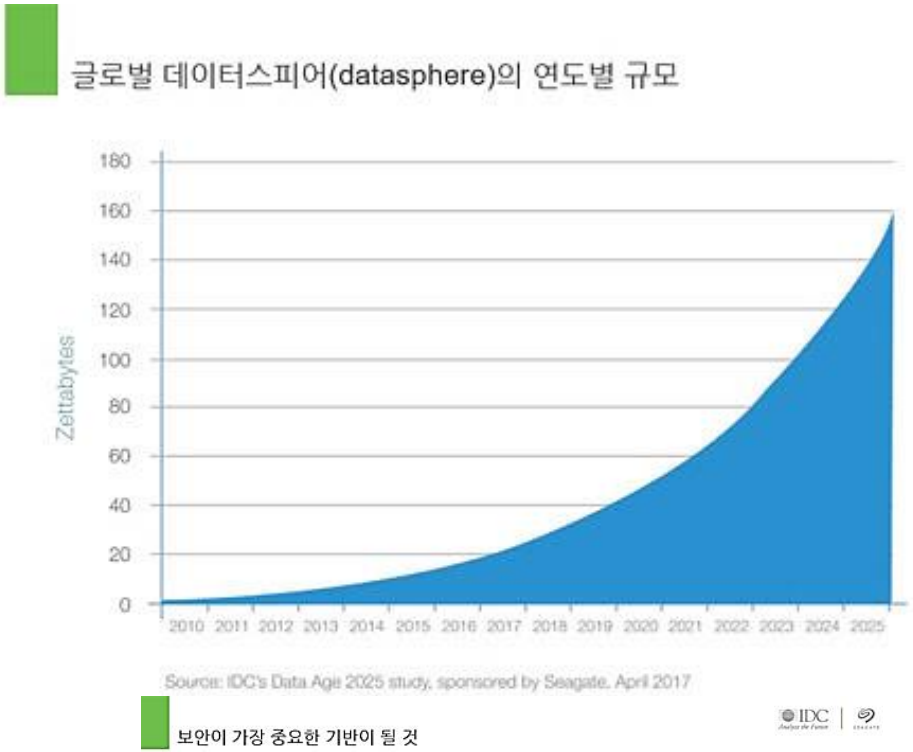
IoT 분야별 보안 위협 시나리오

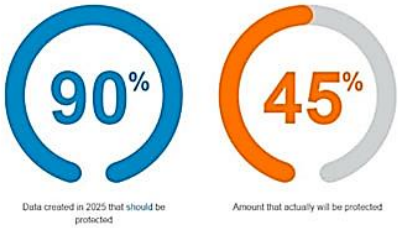
분야	보안 취약성 및 공격 유형
CCTV	CCTV에 탑재된 카메라 해킹, 사생활 영상 추출
스마트 가전	로봇청소기 취약점 해킹, 탑재된 카메라로 실시간 영상 유출
홈	홈 IoT를 해킹, 도어락 해킹, 전력량 해킹, 가스락 해킹, 물 누수, 전등 해킹
공장	기계 오작동, 전력량 해킹, 물 누수, 관제 해킹(CCTV 등)
공유기	공유키 해킹, 악성코드를 넣어 DDoS 공격 창구로 활용
교통	도로차량 감지기술 내 결함, 센서를 가장해 교통관리 시스템에 위변조 데이터 전송
의료기기	인슐린 펌프 조작 해킹, 치명적 복용량 주입
IoT 제조사	불법복제, 유통으로 매출 저하 및 회사 이미지 실추
인명사고 유발	오작동, 악의적 조작으로 신체적/정신적 피해 유발, 법적 책임 문제 발생 및 회사 이미지 실
디바이스 등	Worm, Virus, 기밀성/무결성 공격, 비인가된 접근, 비인가된 I/O 접근, 설정 오류 및 실수, 복제 공격, 보호되지 않는 펌웨어
통신/네트워크	DoS, DDoS의 경유지로 악용, 방화벽의 부적절한 사용, 프로토콜 보안 취약성
플랫폼,응용서비스	패치안된 시스템 OS, OS 보안 취약성, Anti-Virus SW의 무분별한 사용, 부적절한 시스템 Log 기록, 프라이버시 침해
응용서비스 무단이용	비인가된 서비스 접근, 비인가된 사용자의 접근, 안전하지 않은 패스워드 사용, 서비스 인프라 구축 및 운용 비용 증가

출처 : KISA

1. IoT 보안

제목 : 2025년 전세계 데이터 163ZB로 10배 증가, 데이터 중요성 강조





기존에는 데이터 생성 주체가 일반 소비자였다면 앞으로는 기업에 의해 머신 투 머신(M to M)이나 **IoT 분야에서 생성되는 데이터가 훨씬 클 것으로 전망**했다. 특히 그 동안에는 일반 소비자들이 사용하는 카메라나 스마트폰, PC 등이 데이터를 생성하는 엔드포인트(Endpoint)였고, ...

예를 들어 **자율주행차는 각종 센서에서 수집되는 주변의 정보를 빠르게 분석해 정확한 의사 결정**을 내려야 하는데 클라우드까지 데이터를 보내서 이를 처리해 다시 받기까지는 시간이 너무 걸린다.

데이터 폭증과 함께 보안의 중요성도 언급했다. **지금 생성되는 데이터의 약 90% 정도는 보안을 필요로 하는 것이지만 실제로 보안이 제공되는 데이터는 45% 정도로 절반에도 미치지 못한다며** 씨게이트는 데이터 센터를 위한 스토리지 솔루션 암호화, HDD 무단 제거시 데이터 삭제, 제품 생산부터 데이터 센터까지 이동 중 해킹 방지, 엔드포인트 취약점에 대비한 솔루션 등 다양한 방법으로 보안을 강화하고 있다고 밝혔다.

보드나라 : www.bodnara.co.kr

1. IoT 보안

스마트 팩토리 보안의 중요성

2. AI에 대한 이해

2016년에 진행 된 알파고와 이세돌의 대국/ 주산 3~4단과 회계 정보 시스템

2. AI에 대한 이해

주산 3~4단과 컴퓨터/회계 정보 시스템

2. AI에 대한 이해

▪ AI의 시작

✓ **1943년 논리학자인 월터 피츠와 신경외과의인 워렌 맥컬럭의 논문**에서 최초의 인간 두뇌에 관한 모델이 등장

✓ **1950년 영국의 수학자 앨런 튜링의 논문**에서 '**생각하는 기계**'에 대해 기술

✓ **1956**년 존 매카시 교수가 '다트머스 AI 컨퍼런스'를 개최하면서 초청장 문구에 '**AI**'라는 용어를 처음으로 사용

▪ AI의 발전

✓ 1950년대의 인공지능 연구는 크게 **기호주의와 연결주의의 두 분야로 전개**

✓ 기호주의: 인간의 지능과 지식을 기호화해 매뉴얼화하는 접근법

✓ 연결주의: 1943년 월터 피츠와 워렌 맥컬럭이 연구한 뇌 신경 네트워크의 재현을 목표로 하는 접근법

✓ 퍼셉트론은 인간의 사진을 대상으로 남자와 여자를 구별해내면서 뉴욕 타임즈에 등장

✓ **퍼셉트론: 인공 신경망(딥 러닝)의 기본이 되는 알고리즘**

▪ AI의 빙하기

✓ 마빈 민스키는 제자 시모어 페퍼트와 퍼셉트론의 한계를 수학적으로 증명

✓ 퍼셉트론이 무너지고 설상가 상으로 2년 뒤인 **1971년에 로젠블랫이 사망하며 신경망 열기가 급격히 냉각**

✓ 로젠블랫의 퍼셉트론으로는 **XOR 같은 비선형 문제는 해결할 수 없음**

▪ 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

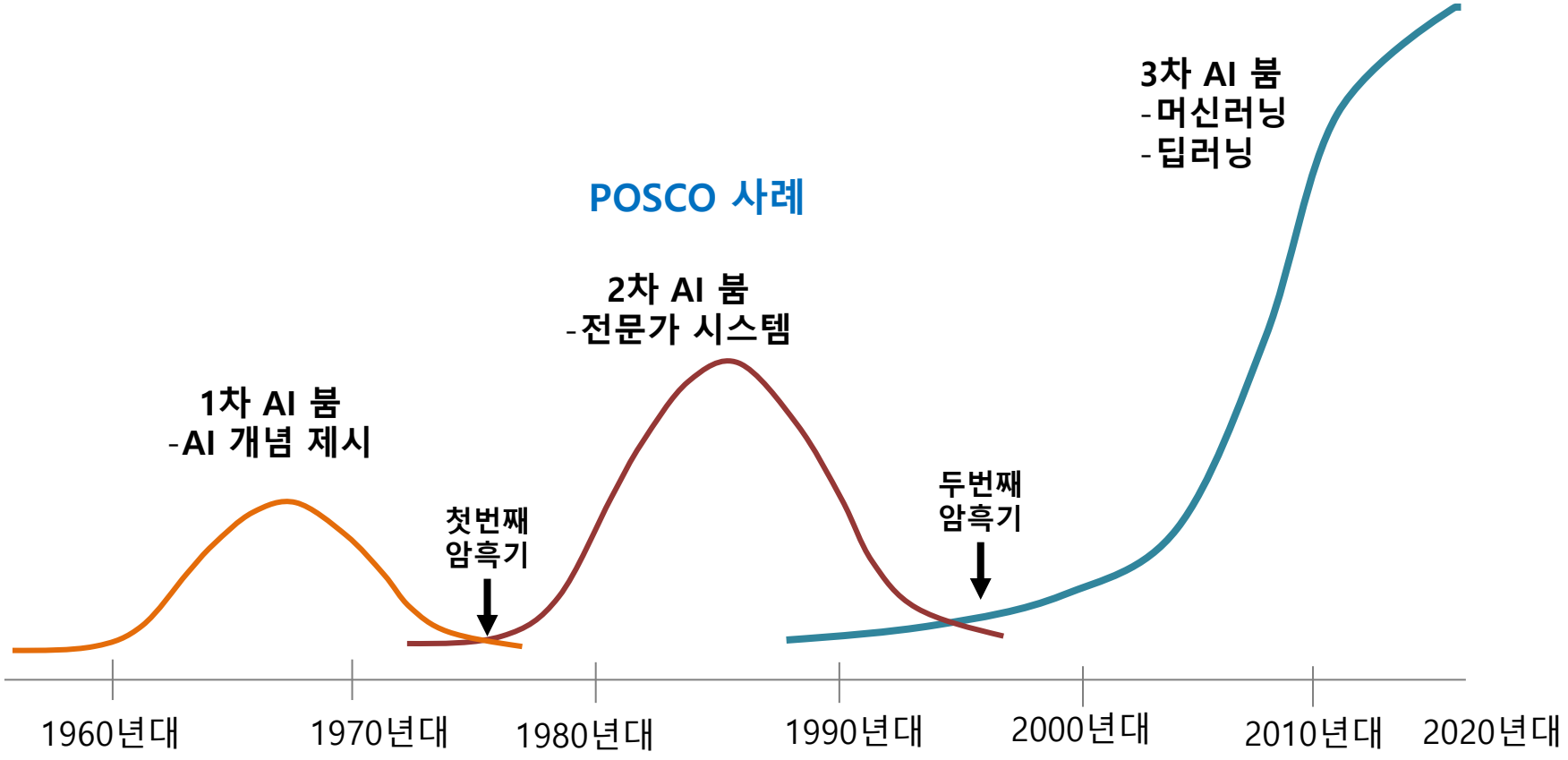
▪ 딥 러닝의 시작

✓ 제프리 힌튼은 가중치의 **초깃값을 제대로 설정한다면 깊은 신경망을 통한 학습이 가능하다는 것을 밝혀 냄.**

✓ **인공 신경망**이라는 단어가 들어간 논문을 학회에 투고하면 제목만 보고 거절당하거나 사람들의 관심을 끌지 못하기 때문에 제프리 힌튼은 **이 논문에 deep을 붙인 DNN이라는 용어를 사용**

2. AI에 대한 이해

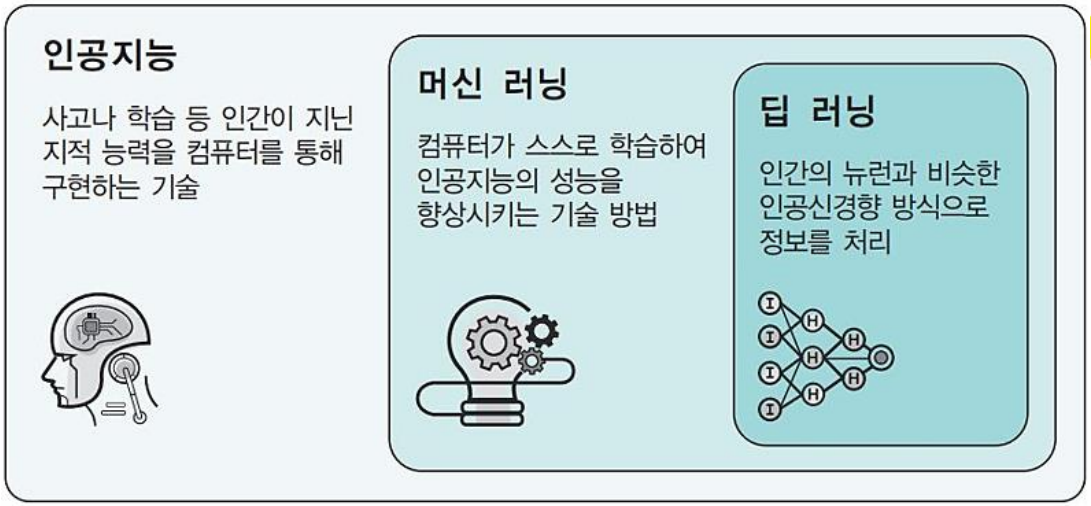
인공지능(AI) 발전 과정

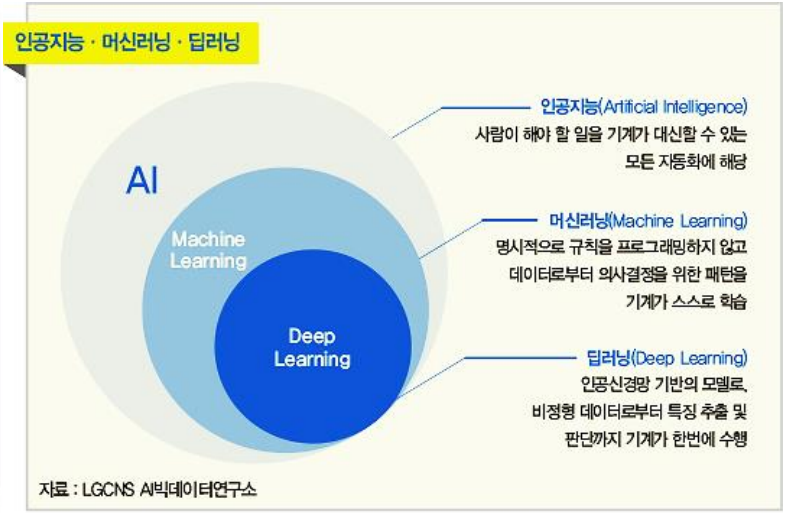


자료 : 마쓰오 유타카(2015년)

AWS(Amazon Web Service)의 AI 정의

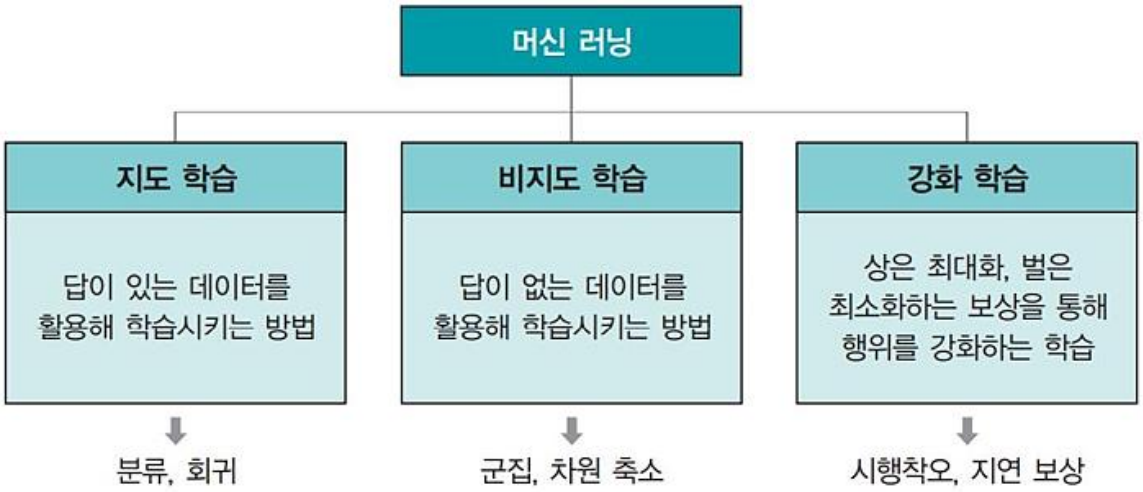
Artificial Intelligence (AI) is the field of computer science dedicated to solving cognitive problems commonly associated with human intelligence, such as **learning, problem solving, and pattern recognition**.





머신 러닝의 분류

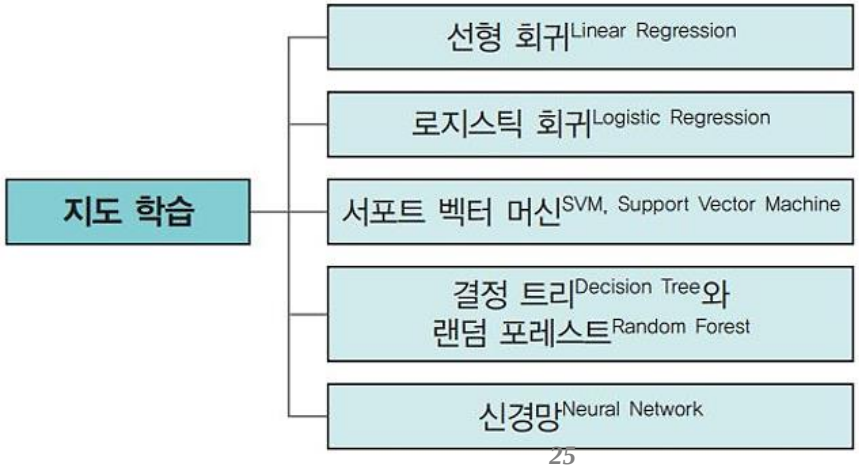
- 머신 러닝 역시 상당 부분 통계학적인 기술을 포함
  - ✓ **지도 학습** : 분류나 회귀에 사용
  - ✓ **비지도 학습** : 군집에 사용
  - ✓ **강화 학습** : 환경에서 취하는 행동에 대한 보상을 이용하여 학습을 진행



머신 러닝의 분류(지도 학습)

답이 있는 데이터를 활용해 학습시키는 방법, 입력 값(X) 이 주어지면 입력 값에 대한 라벨(Y)을 주어 학습

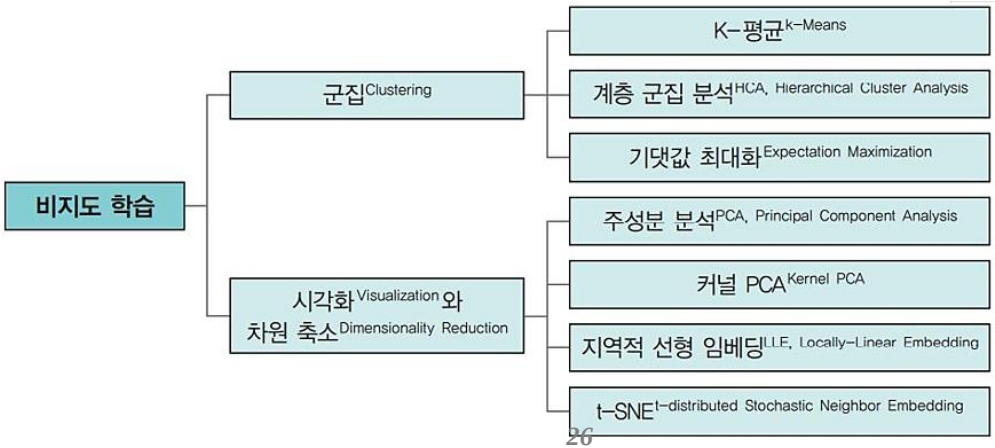
- 분류
  - ✓ 주어진 데이터를 정해진 레이블(범주)에 따라 나누는 것
  - ✓ 범주가 2개 이면 이진 분류 , 3개 이상이면 다중 클래스 분류
- 회귀
  - ✓ 어떤 데이터들의 특징을 토대로 값을 예측하는 것
  - ✓ 결과 값은 실수 값을 가질 수 있음



머신 러닝의 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 변수 중에 유의미한 변수들을 식별하여 개수를 줄이는 작업
  - ✓ 일부 변수가 중복되거나 작업과 아무 관련이 없는 경우가 많기 때문
  - ✓ 변수를 줄이면 잠재되어 있는 진정한 관계를 도출하기가 용이

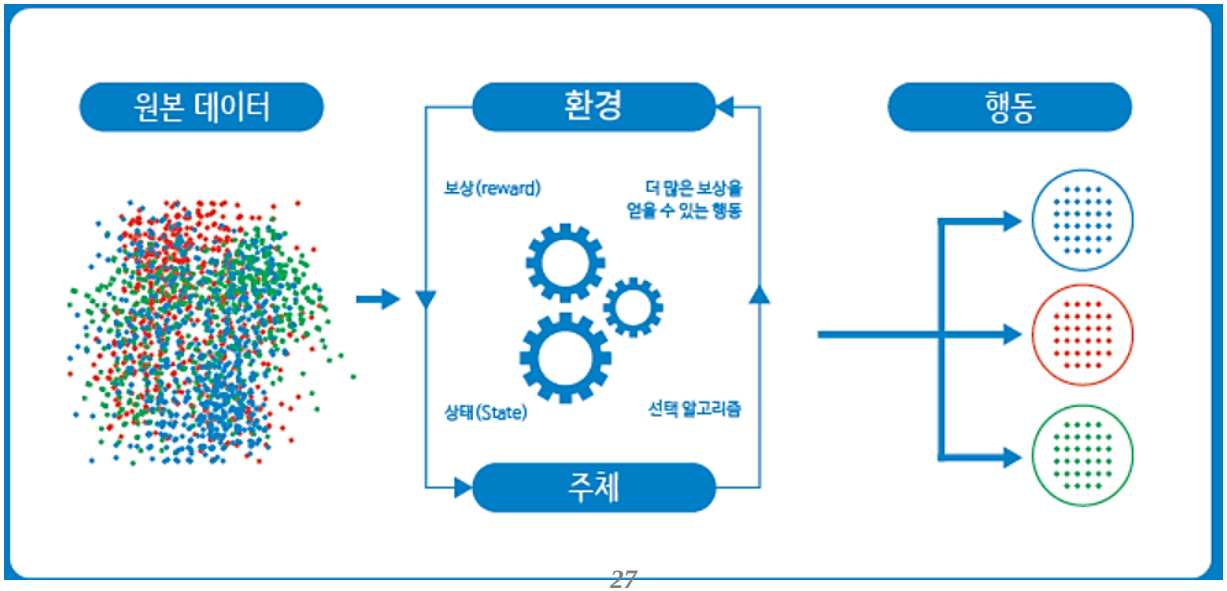


2. AI에 대한 이해

머신 러닝의 분류(강화 학습)

- 지도·비지도 학습과는 다른 종류의 알고리즘
- 학습하는 시스템을 '에이전트'라고 부르며, 환경을 관찰해서 행위를 수행하고 보상 받음
- 시간이 지나면서 **가장 큰 보상을 얻기 위해 '정책'이는 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

강화 학습은 주체(agent)가 환경으로부터 보상을 받음으로써 학습하기 때문에 지도 학습과 유사해 보이지만, 사람으로부터 학습을 받는 것이 아니라 변화되는 환경으로부터 보상을 받아 학습한다는 점에서 차이를 보임.



3. AI의 취약점 유형과 대안

3. AI의 취약점 유형과 대안

- 해커의 공격 일반적인 공격 : 네트워크 및 시스템의 취약점을 이용한 공격
- AI에 대한 공격 : Data를 이용한 공격
- AI의 취약 유형
  - ✓ **Data 변조 공격**
  - ✓ **악의적 Data 주입 공격**
  - ✓ **Data 축출 공격**

2016년 미국의 한 쇼핑센터에서 경비 역할을 하던 로봇이 갑자기 오류를 일으켜 16개월 된 아이를 넘어뜨리고 그대로 지나가 아이의 다리를 다치게 하는 사고가 일어났다. 같은 해 중국에서도 한 IT전시회에서 시연 중이던 교육용 로봇이 갑자기 전시장 유리를 깨뜨리고 이 유리 파편으로 인해 방문객이 부상을 입는 사고가 일어났다. 2018년에는 미국 뉴저지에 있는 아마존 물류센터에서 일하던 로봇이 갑자기 오류를 일으켜 곰 퇴치 스프레이 통을 찢어 버렸고, 유독물질이 유출돼 무려 24명의 직원이 병원에 입원한 사고도 일어났다. 이 사고 이후 아마존 직원 노동조합에서는 아마존의 로봇이 인간 근로자에게 끼치는 위험에 대해 경고하는 성명을 발표했다.

3. AI의 취약점 유형과 대안

정 보 보 호 학 회 지
제 30 권 제2호, 2020. 4

자율 주행 자동차 보안 위협 및 기술 동향

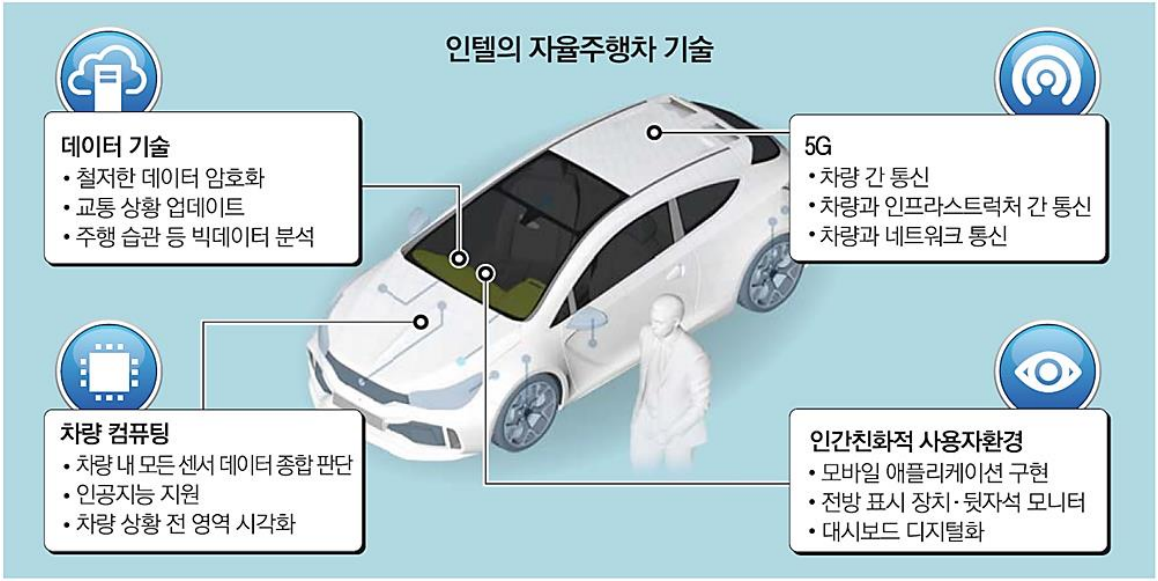
권 순 홍*, 이 종 혁*

요 약

IT 기술을 차량에 적용하여 사람의 조작 없이 차량 스스로 운행하는 자율 주행 자동차에 대한 연구가 활발하게 진행되고 있으며 상용화 및 대중화에 집중하고 있는 추세이다. 자율 주행 자동차의 경우, 보안 취약점을 통한 공격을 통해 오류를 발생시킬 경우, 운전자 또는 보행자에게 직접적인 해를 끼칠 수 있어 보안 취약점 및 보안 기술에 대한 연구는 자율 주행 자동차 상용화 및 대중화에 있어 핵심적인 부분이라고 할 수 있다. 본 논문에서는 현재 자율 주행 자동차의 기술 단계와 작동 원리에 대해 설명하며, 자율 주행 자동차의 보안 위협 요소를 살펴보고, 보안 위협으로부터 운전자 또는 보행자를 보호할 수 있는 보안 기술 현황에 대해 설명한다.

3.1. 자율 주행 자동차 보안 위협 동향

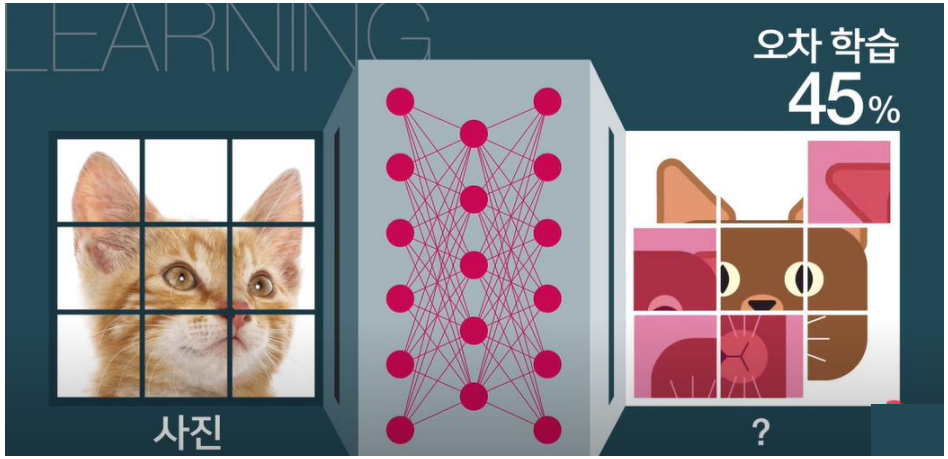
기존 차량과 비교하여 자율 주행 차량의 경우, 외부 네트워크를 통해 차량 대 차량 및 차량 대 인프라간의 통신을 수행한다. 이처럼 외부에 노출되는 외부 네트워크를 통해 공격자는 보안 취약점을 이용하여 공격을 수행함으로써 사람에게 직접적인 피해를 입힐 수 있다. 또한, 자율 주행 차량은 센싱을 통해 방대한 양의 정보를 수집하고 이를 딥러닝 알고리즘을 통해 처리한다. 자율 주행 차량이 딥러닝을 통해 방대한 양의 정보를 처리할 때, 공격자는 정상적이지 않은 데이터를 주입함으로써 예상치 못한 상황을 야기시킬 수 있다. 자율 주행 자동차가 상용화 및 대중화되기 위해서는 이와 같은 문제가 발생하면 안되므로 철저한 하드웨어 및 소프트웨어 테스트가 요구된다.

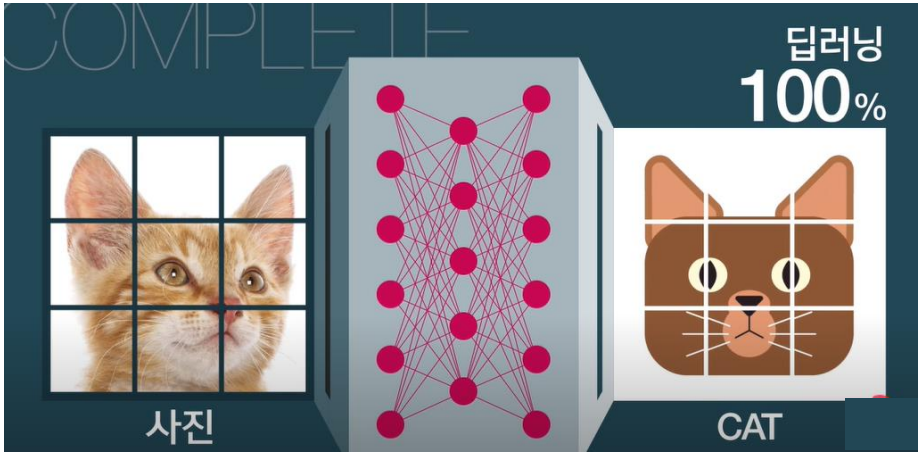


3. AI의 취약점 유형과 대안

회피 공격(Evasion Attack)

- 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
- 데이터가 변조되었다면 변조 공격을 학습 데이터에 포함해 훈련시키는 방법으로 대응할 수 있음

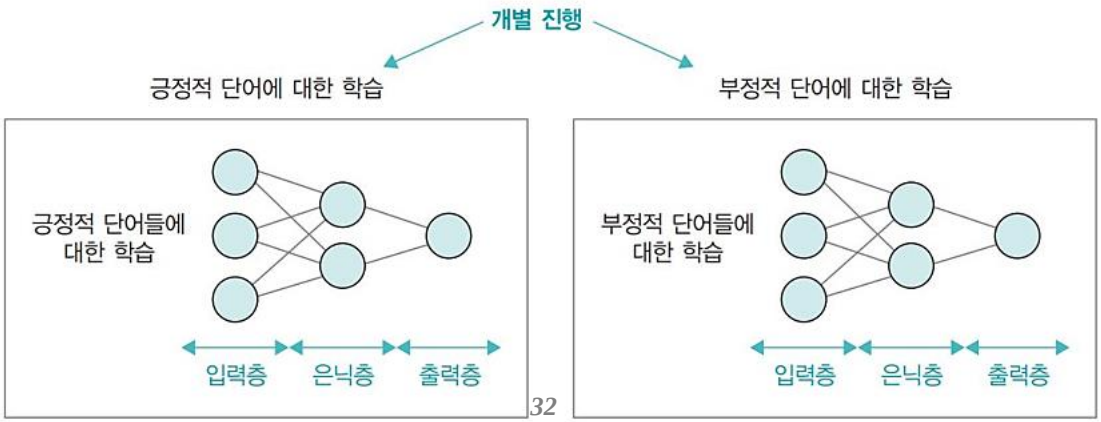




3. AI의 취약점 유형과 대안

중독 공격(Poisoning Attack)

- **악의적인 데이터를 이용해 인공지능 시스템이 오작동을 일으키도록 하는 공격**
- 예) 스캐터랩의 '이루다'는 일부 사용자들이 이루다에게 욕설, 인종 차별 및 성 차별 발언, 정치적 발언 등 악의적인 발언들을 훈련시키면서 정상적인 서비스가 불가능해짐
- **<u>부정적인 데이터에 대한 사전 학습으로 대응할 수 있음</u>**
- 사전 학습된 단어들과 비교하여 추가 학습이 불가 하거나 답변을 우회하도록 프로그램을 설계하는 방법도 고려



3. AI의 취약점 유형과 대안

중독 공격(Poisoning Attack)

스캐터랩은 지난 1년간 '보완작업'에 돌입

- 전반적인 AI 윤리를 점검하고, 내부 기획자·리서처·엔지니어 등과 논의를 거쳐 윤리 준칙을 수립했다.
- 개인정보보호를 강화할 수 있도록 딥러닝 알고리즘이 생성한 문장으로 답변할 수 있도록 했다.
- "대화 시 특정 단어·문맥을 탐지해 선정적이거나 공격적, 또는 편향적 문장에 대응할 수 있도록 '**어뷰징** 탐지 모델'을 접목

출처 : https://www.bloter.net/newsView/blt202112210023

3. AI의 취약점 유형과 대안

어뷰징(Abusing,남용,오용)

- 남용, 악용, 학대, 욕설 등을 뜻하는 단어인 Abuse에서 파생된 단어로 주로 온라인 게임에서 버그, 핵 등의 불법 프로그램, 타인 계정 도용, 다중 계정 접속 등을 통해 부당한 이득을 챙기는 행위.

- 승부조작도 어뷰징의 일종이라 볼 수 있다. 다만 승부조작을 일컫는 정확한 영어 표현은 따로 존재하며, **매치 픽싱(Match Fixing)**이라 함.

- 버그나 글리치를 악용하는 행위는 어뷰징 중에서도 Exploit에 해당.

- **글리치**(영어: glitch)는 시스템의 일시적인 오류를 일컫는 말

3. AI의 취약점 유형과 대안

전도 공격(Inversion Attack)

- **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

4. AI를 이용한 보안

사이버 보안 세계에서는 속도가 피해 발생 여부를 좌우할 결정적인 요소이다. 교묘한 공격 수법으로 단 20분 이내로 중대한 데이터 자산 보안을 공격할 수도 있다. 사이버 공격에 첨단 기술을 동원한 사례가 증가한 탓에 인간이 악성 활동을 재빨리 감지하는 것이 어려워졌다.

더 심각한 점은 마이크로소프트 보안·규정준수·신원 확인 부사장인 앤 존슨(Ann Johnson)은 미국 온라인 IT 매체 벤처비트와의 인터뷰에서 "인간의 사이버 공격 감지가 갈수록 어려워지는 상황에서 인공지능(AI) 기술을 활용해 신속한 보안 솔루션을 최대한 제공해야 한다"라고 주장했다. 사이버 보안 신속 대응에서 AI가 중요한 이유는 무엇일까?

① 사이버 공격 조기 차단
② 보호된 데이터 기록 및 분류
③ 제로 트러스트 아키텍처 구축(zero-trust architecture)
④ 사소한 업무 자동화

출처 : 코딩월드뉴스 (https://www.codingworldnews.com)

4. AI를 이용한 보안

스팸 메일 개요

- 일반적으로 스팸 메일은 프로그램을 이용하여 **불특정 다수의 사용자에게 송신.**
  - ✓ 송신되는 **스팸 메일은 주로 음란물 사이트, 도박 사이트 , 불법 사이트** 등을 광고하는 광고성 메일이 대부분.
  - ✓ 근래에는 광고성 메일에 사용자 정보를 탈취하는 악성링크 및 첨부파일이 다수 삽입되어 APT 공격의 주요 수단으로 활용.
- 대부분의 **랜섬웨어** 또한 메일을 통해 유포되었다. 이러한 악성 메일도 프로그램을 통해 송신되기 때문에 **스팸 메일과 동일하게 고유의 패턴이 존재**
- **악성 메일에 대응하는 방법 :** 두 가지.
  - ① 사용자들이 '의심스러운 메일의 첨부파일 및 URL 접근금지'와 같이 조심하는 것이고
  - ② **패턴을 활용한** '**스팸 메일 차단 솔루션**'을 이용하여 악성 메일을 차단하는 방법이다. 일반적으로 스팸 메일 차단 솔루션을 통해서 스팸 메일의 70~90%는 차단이 가능

4. AI를 이용한 보안

- **나이브 베이즈 분류기(Naive Bayes Classfier)**
  - ✓ 스팸 메일 솔루션에 일반적으로 적용된 알고리즘
  - ✓ 인공 신경망 알고리즘에는 속하지 않지만 머신 러닝의 주요 알고리즘으로 분류되어 있고 준수한 성능을 보임

- **베이브의 정리(Bayes' theorem)**
  - ✓ 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

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

4. AI를 이용한 보안

베이즈의 정리(Bayes' theorem)

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

건빵 2봉지를 샀다. 그래서 별사탕도 2봉지다. 첫 번째 봉지에는 하얀 별사탕이 1개, 분홍 별사탕이 3개 들었고, 두 번째 봉지에는 각각 2개씩 들었다. 두 봉지의 별사탕을 하나의 접시에 담고, 눈을 감은 채 별사탕 하나를 집어 들었다. 눈을 뜨고 집어 든 별사탕을 지그시 살펴보니 분홍별사탕이다. **이 별사탕이 첫 번째 봉지에서 나왔을 확률은?**

P(첫번째봉지|분홍별사탕) = P(분홍별사탕|첫번째봉지)P(첫번째봉지)/P(분홍별사탕)

P(분홍별사탕|첫번째봉지) = 3/4
P(첫번째봉지) = 4/8
P(분홍별사탕) = 5/8

(3/4) * (4/8) / (5/8) = 3/5 = 60%

4. AI를 이용한 보안

나이브 베이즈 분류기(Naive Bayes Classfier)

- 입력 텍스트가 정상 메일인지 스팸 메일인지 구분하기 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)**

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

- **P(정상 메일|텍스트)가 P(스팸 메일|텍스트)보다 크다면 정상 메일이라고** 볼 수 있으며, 그 반대라면 스팸 메일
- 메일 본문에 있는 단어가 3개(w1, w2, w3), 나이브 베이즈 분류기의 정상 메일일 확률과 스팸 메일일 확률을 구하는 식은 (두식에서 동일하게 존재하는 P(텍스트)를 제거)
  - ✓ P(정상 메일|텍스트) = P(w1|정상 메일) * P(w2|정상 메일) * P(w3|정상 메일) * P(정상 메일)
  - ✓ P(스팸 메일|텍스트) = P(w1|스팸 메일) * P(w2|스팸 메일) * P(w3|스팸 메일) * P(스팸 메일)

4. AI를 이용한 보안

나이브 베이즈 분류기: 예시(get free lottery)

구분	단어	분류	구분	단어	분류
1	chance free lottery	스팸 메일	4	free to contact me	정상 메일
2	get free ticket	스팸 메일	5	you won award	정상 메일
3	get free scholarship	정상 메일	6	you ticket lottery	스팸 메일

- 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 |입력 텍스트) = 1/10 * 2/10 * 0/10 * 0.5 = 0
- P(스팸 메일|입력 텍스트) = 1/9 * 2/9 * 2/9 * 0.5 = 0.28

결과적으로 보면 P(정상 메일|텍스트)=0 보다 P(스팸 메일|텍스트)=0.28 이 큼으로, **get free lottery는 스팸 메일로 분류**

4. AI를 이용한 보안

Google BERT

자연 언어 처리(NLP)를 위한 종래의 방법을 넘은 성능을 발휘한다. BERT는 자연언어 처리 태스크를 교사 없이 **양방향으로 사전학습하는 첫 시스템**이다.

구글이 공개한 인공지능(AI) 언어모델 'BERT(이하 버트, Bidirectional Encoder Representations from Transformers)'는 일부 성능 평가에서 인간보다 더 높은 정확도를 보이며 2018년 말 현재, 자연 언어 처리(NLP) AI의 최첨단 딥러닝 모델

BERT를 활용한 스팸 메일 솔루션은 90% 후반대의 정확도를 보임.

4. AI를 이용한 보안

네트워크 침입 탐지 기술은 속성상 네트워크 트래픽을 분석하여 침입탐지를 확인해야 하므로 효율적 탐지에 대한 요구가 지속적으로 발생

- **전문가 시스템**
  - ✓ ***특정 응용 분야 전문가의 지식 및 능력을 체계적으로 잘 조직하여 컴퓨터 시스템에 입력해 해당 분야의 비전문가 라도 전문가에 상응하는 능력을 발휘할 수 있도록 쉽고 빠르게 도움을 주는 시스템***
  - ✓ 전문가 시스템은 앞서 AI의 역사에서 언급한, 논리적인 체계로 문제를 푸는 기호주의 (Symbolism) 의 한 분야에 속함

- **새로운 접근은 네트워크 트래픽의 패턴 분석 : 머신 러닝 모델 적용**
  - ✓ **오탐률(Error Rate)이 문제**
  - ✓ 거짓 양성**Type 1** 에러라고 하는데, 실제로는 공격이 아닌데 공격이라고 탐지하는 것
  - ✓ 기존의 전문가 시스템은 이와 반대로 **거짓 음성 Type 2 에러가 상대적으로 높음**
  - ✓ Type 2 에러는 공격을 받았으나 이를 탐지하지 못하는 것

4. AI를 이용한 보안

- **기호 주의(Symbolism) AI**
  - ✓ 컴퓨터 작동 방식에 맞게 기호와 규칙을 사용하는 규칙 기반(Rule-based) 인공지능으로 오래전부터 지금까지 지속적으로 사용되고 있는 방식이다.
  - ✓ 컴퓨터 작동 방식으로 인공지능을 구현할 수 있다는 논리이며, 논리적으로 설명 가능한 문제를 다룬다. A=B이고, B=C일 때, A=C이다로 귀결시킬 수 있는 인공지능이다.
- **연결 주의(Connection) AI**
  - ✓ 뉴런의 연결을 모방한 정보처리 과정을 사용하는 **신경망(Neural Network) 기반 인공지능이다**.
  - ✓ 뇌처럼 하나의 개념이 여러 곳에 흩어져서 표현되며, 논리적으로 설명하기 어려운 문제를 다룬다. 특정 세기 이상의 값에 도달하면 0과 1로 정보를 처리한다는 개념이다.

4. AI를 이용한 보안

- **기존에는 악성 코드를 탐지하기 위해 악성 코드의 일부분을 매칭해보거나 특정 부분의 해시 값을 생성하여 비교해보는 등의 방법이 사용**
- 최근에는 이런 **탐지 방법을 회피한 고도화된 방식**으로 다양한 신·변종 악성 코드가 나타나 탐지가 어려움
- 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
  - ✓ API Call
  - ✓ Runtime Log
  - ✓ 시스템 지원
  - ✓ 네트워크

4. AI를 이용한 보안

안드로이드 응용프그램의 행위 기반 분석에 사용되는 특징

특징	설명
API Call	API는 응용프로그램의 수행작업을 파악할 수 있으며, 간접적으로 행위의 의도를 추론할 수 있음
Runtime Log	입출력 네트워크 데이터,파일의 읽기 및 쓰기작업, DexClassLoader를 통한 서비스 시작, 로드된 클래스, 네트워크를 통한 정보 유출, 파일 및 SMS 정보를 포함
시스템 자원	네트워크 액세스, 연락처, SMS송수신 기능 같은 정보를 포함. 응용프로그램이 리소스를 사용할 경우 **개발자는 컴파일할 때 해당 리소스와 관련된 권한 요청**
네트워크	악성 코드가 발생 시킨 패킷의 시작과 종료 시간, 송.수신 패킷의 플로,IP 주소, Port번 등의 정보를 통한 악성 코드 탐지.

4. AI를 이용한 보안

- 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 애저 클라우드 환경에서 분석하여 주유소와 관련한 위험 요인을 탐지

Azure란?

Azure 클라우드 플랫폼은 새로운 솔루션을 구현하여 현재의 문제를 해결하고 미래로 나아가는 데 도움이 되도록 설계된 200개가 넘는 제품 및 클라우드 서비스입니다.

4. AI를 이용한 보안

국가직무능력표준(NCS)

< 이전화면

20. 정보통신

중분류	소분류	세분류	능력단위
01. 정보기술	01. 정보기술전략·계획	01. 정보보호관리·운영	01. 지능형영상정보처리시스템 요구사항 분석
02. 통신기술	02. 정보기술개발	02. 정보보호진단·분석	02. 지능형영상정보처리시스템 설계
03. 방송기술	03. 정보기술운영	03. 보안사고분석대응	03. 지능형영상정보처리 알고리즘 분석
	04. 정보기술관리	04. 정보보호암호·인증	04. 지능형영상정보처리 알고리즘 설계
	05. 정보기술영업	05. 지능형영상정보처리	
	06. 정보보호	06. 생체인식(바이오인식)	
	07. 인공지능	07. 개인정보보호	

Thank you