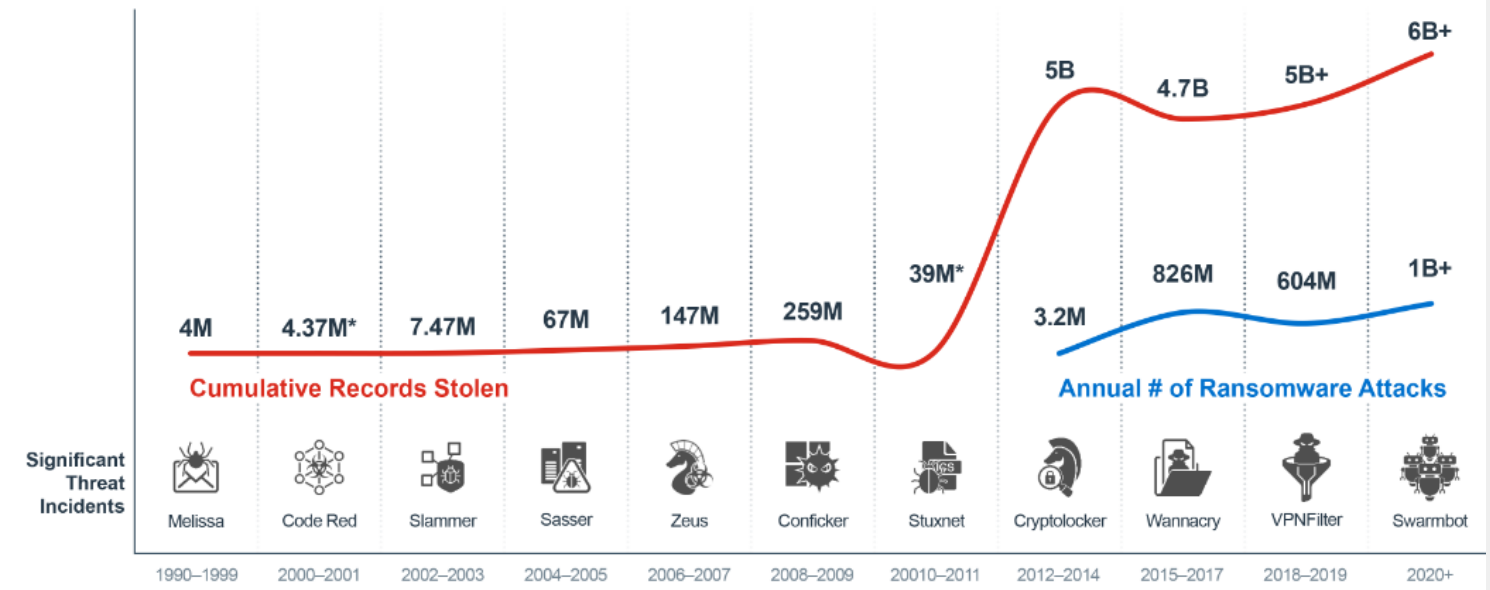
# AI 기반 악성코드 탐지

조서윤

# 최근 20년 간의 위협 동향



## 1 : Signature Based

- 정적 분석 수행
- 빠른  탐지 시간
- 대표 솔루션 : Antivirus
- 시그니처 없는 경우 미탐 (변종·신종 악성코드 취약)

## 2 : ATP(Advanced Threat Protection) Toolkit

- 정적 및 동적 분석 수행
- 진화된 악성코드 탐지 가능
- 상세한 악성파일 분석 보고서 제공
- 느린 분석 시간 -> 느린 대응
- 대표 솔루션 : Sandbox
- 자동화된 악성코드 동적 분석 수행
- 대용량 망에서 높은 비용 발생

## 3 : Artificial Intelligence

- 정적 분석 수행
- 대표 솔루션 : FortiAI
- 딥러닝 기반 악성코드 특징 분석
- 사전 학습 데이터셋을 통한 정상/비정상 파일 분류
- 설치된 환경에 대한 지속 학습을 통해 탐지력 점차 강화
- 빠른 분석 시간 -> 빠른 대응
- 기존 운영중인 Sandbox 연계를 통해 부하 감소

# '사이버보안 AI·빅데이터 챌린지 2022'에서 9개 팀 수상

**K스피릿** | 입력 2022.12.01 15:43 | 업데이트 2022.12.01 15:48

정유철 기자

'사이버보안 AI·빅데이터 챌린지 2022'에서 null@root(악성코드 분야)팀과 0xFFFA(침해사고 분야)팀이 각각 대상을 수상했다.

과학기술정보통신부(장관 이종호, 이하 '과기정통부')와 한국인터넷진흥원(원장 이원태, 이하 KISA)은 사이버보안 분야 AI·빅데이터 활용 대국민 관심도 제고를 위해 개최한 '사이버보안 AI·빅데이터 챌린지 2022' 시상식과 우수성과 공유회를 12월 1일(목) 서울 잠실 롯데타워SKY31 오디토리움에서 개최했다.

올해 챌린지에는 △악성코드 △침해사고를 주제로 기술경연 2개 분야, △AI활용 아이디어 공모 1개 분야 총 3개 분야에 정보보호 산업계, 학계, 연구기관 등 모두 85개 팀, 222명이 참가 신청서를 접수하였다.

이 가운데 총 21개 팀이 본선 진출 하였으며, 본선 진출 팀 대상으로 실시된 기술경연, 발표평가에서 우수 평가를 받은 9개 팀이 최종 수상자로 결정되었다.

수상자는 악성코드 분야에서 null@root팀이 대상(과기정통부장관賞), Xransformer팀이 최우수상(한국인터넷진흥원장賞), 멀웨어가뭐라고생각하세요팀이 우수상(안랩사장賞)을 받았다.

침해사고 분야에서는 0xFFFA팀이 대상(과기정통부장관賞), SSAPGOSU팀이 최우수상(한국인터넷진흥원장賞), 소떡말고네떡팀이 우수상(이글루코퍼레이션사장賞)을 받았다. 아이디어 공모 분야에서는

# "넥슨코리아, 악성코드 분석률 50%→100%"...AI 데이터셋 구축으로 민간 사이버보안 대응력 높였다

A 윤명주 기자 | 입력 2022.07.21 15:40 | 댓글 0 | 좋아요 0

'AI 활용 사이버보안 대응체계 고도화' 지원 나선다
사이버보안 분야 특화 AI 학습 데이터셋 8억 건 구축
데이터셋 구축 실증 결과 등 추진성과·활용사례 공유
"양질의 사이버보안 AI 데이터셋을 확대 구축할 계획"
민간분야 사이버보안 대응체계 고도화에 기여할 것


21일 개최된 '2021년 사이버보안 AI 데이터셋 구축 사업성과 공유회'에서 김장성 과학기술정보통신부 정보보호네트워크정책관이 인사말을 하고 있다. (사진=한국인터넷진흥원 유튜브 채널 캡처).

과학기술정보통신부와 한국인터넷진흥원(KISA)이 지난해 사이버보안 인공지능(AI) 데이터셋(악성코드·침해사고 분야) 구축 추진 성과를 공유하고 우수 활용사례를 소개하는 자리를 가졌다. 21일 서울시 중구 더플라자 호텔 메이플홀에서 열린 '사이버보안 AI 데이터셋 구축 성과 공유회'를 통해 민간분야 AI 기반 보안 대응체계 고도화를 위한 심도 있는 논의의 장을 마련한 것.

AIS 2022

2022 인공지능 보안 컨퍼런스

데일리시큐


AI 기술을 이용한
악성코드 탐지 및 방어

누리랩

최원혁 대표이사

개인적인 결론입니다만…

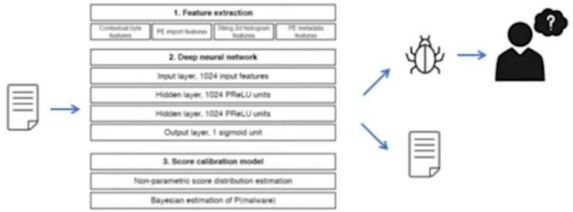
## (1) AI가 탐지 결과에 대한 신뢰성 결여

**시그니처 백신**

- 악성코드 분석 전문가가 분석 후 시그니처 생성
- 악성코드 탐지 시 사용자의 의심이 크지 않음

**AI 백신**

- 빅데이터를 기반으로 학습 결과물을 이용해 악성코드 탐지
- 사용자는 궁금하다 = 어떤 이유로 악성코드로 탐지가 되었나요?
- 악성코드 탐지 이유 설명 불가



AIS 2022
2022 인공지능 보안 컨퍼런스

데일리시큐

AI 기술을 이용한
악성코드 탐지 및 방어

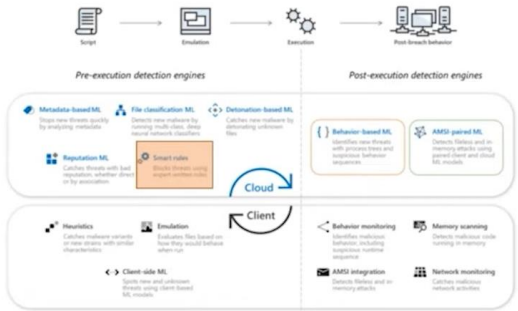
누리랩
**최원혁** 대표이사

개인적인 결론입니다만…

## (2) 신종 악성코드 탐지 불가

AI 백신

- 빅데이터를 기반으로 학습 결과물을 이용해 악성코드 탐지 → 즉, 과거 악성코드를 기반으로 학습
- 신종 악성코드는 이전 사례가 없으므로 탐지 불가
  - 첫 번째 신종 악성코드는 놓일 수 있음
    (이는 시그니처 기반 백신도 마찬가지)
  - 두 번째 악성코드가 나오더라도 학습을 위한
    충분한 데이터가 존재하지 않으므로 학습 불가
  - AI 백신이라도 시그니처 탐지 기법을
    보조 방식으로 사용할 수 밖에 없음



AIS 2022
2022 인공지능 보안 컨퍼런스

데일리시큐

AI 기술을 이용한
악성코드 탐지 및 방어

누리랩
**최원혁** 대표이사

# (4) AI 기법뿐만 아니라 시그니처 기법 도입

AI 기법 / 시그니처 기법을 통해 상호 보완

- 이미 두 방법을 통해 미탐율 감소 확인됨
- 백신을 통해 악성 함수 위치 및 크기 추출
  → yara rule 자동 생성 처리

백신 탐지율 테스트

| 파일 종류 | 시그니처 엑신 | 휴리스틱 엑신 | 시그니처와 의심라벨 결합 |
|---|---|---|---|
| 실행파일 | | | |
| 미탐지율 | 10.1 | 1.1 | 0.5 |
| 문서전체 | | | |
| 미탐지율 | 14.4 | 10.7 | 2.7 |
| 위치찾아 | | | |
| 내탐지율 | 17.0 | 0 | 0 |
| 변종 악성코 | | | |
| 드 미탐지율 | 93.8 | 0 | 0 |

```
sub_00405cf9 ok
sub_00405d02 ok
sub_00405d0b ok
sub_00405d25 ok
sub_00405d41 ok
sub_00405dc8 ok
sub_00405deb ok
sub_00405e14 ok
sub_00405e50 ok
tmpplmvhlvl.__entry malware
sub_00405f22 ok
sub_00405f97 ok
sub_00405fbb ok
sub_00406003 ok
sub_0040609e ok
sub_00406251 ok
sub_004062fb ok
sub_00406358 ok
```

# XAI

설명 가능한 인공지능(Explainable Artificial Intelligence)의 약자로, 인공지능의 행위와 도출한 결과를 사람이 이해할 수 있는 형태로 설명하는 방법

인공지능 시스템의 도출 결과를 이해하지 못하는, 블랙박스 한계를 해결하고자 등장

**과정**

1. 기존 머신러닝 모델에 설명 가능한 기능 추가
2. 머신러닝 모델에 HCI(Human Computer Interaction) 기능 추가
3. XAI를 통한 현재 상황의 개선

**구현 방법 1**

1. 현재 문제를 해결하는 머신러닝 모델을 만든다.
2. 설명가능한 모델을 결합한다.
3. 모델의 결과를 해석하는 인터페이스를 연결한다.
4. 모델의 문제점을 발견하고 개선한다.
5. 모델을 테스트하고 평가하는 파이프라인을 구축한다.

**구현 방법 2**

1. XAI 구현 방법 1을 설명하는 이론적 기반을 마련한다.
2. (1)을 뒷받침할 수 있는 계산 모델(Computational Model)을 만든다.
3. 모델을 검증한다.

# 악성코드 유형 분류를 위한 인공지능 모델 비교 및 분석에 관한 연구

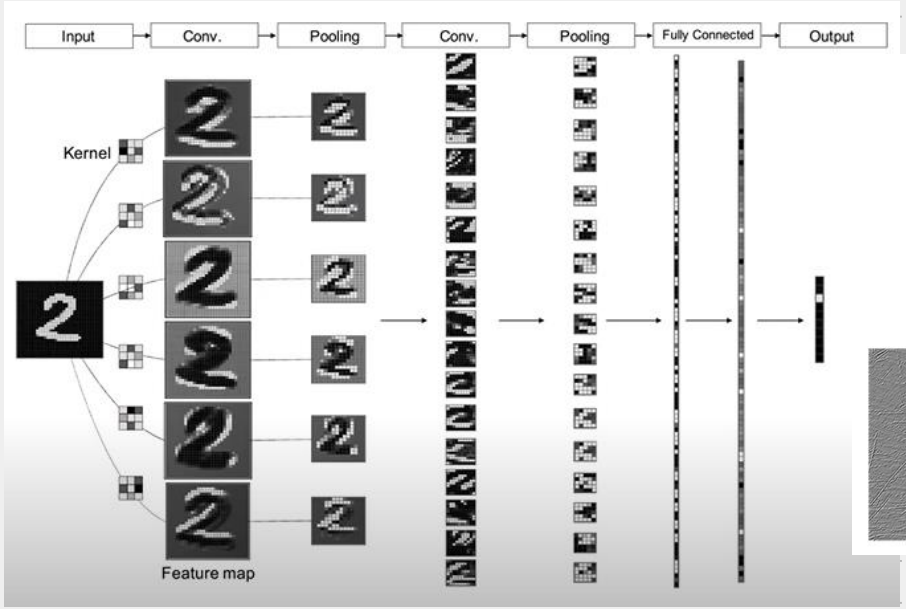
2023 한국컴퓨터종합학술대회 논문집

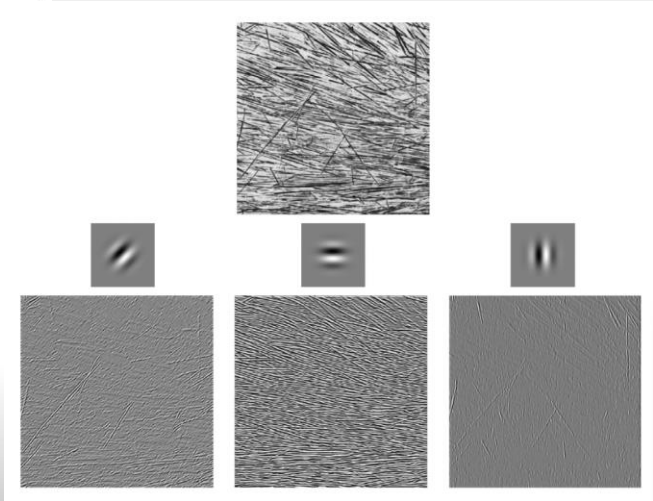
표 2 모델 성능 비교 결과

| Model | Accuracy | F-Score |
|---|---|---|
| k-NN | 0.9461 | 0.9465 |
| Logistic Regression | 0.7498 | 0.7284 |
| Random Forest | 0.9825 | 0.9825 |
| CNN | 0.9313 | 0.9324 |
| LSTM | 0.9696 | 0.9698 |

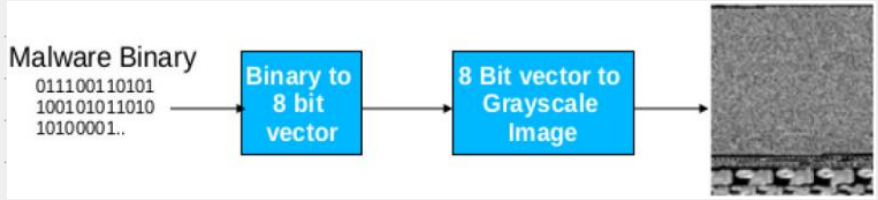
- K-NN : 지도학습 알고리즘, 특정 데이터를 정해진 클래스에 따라 분류
- Logistic Regression : 데이터가 특정 카테고리에 속할지 0과 1사이의 연속적인 확률로 예측하는 회귀 알고리즘
- Random Forest : 오버피팅을 방지하기 위해, 최적의 기준 변수를 랜덤 선택하는 머신러닝 기법
- CNN : 이미지나 영상 데이터를 처리할 때 사용되는 딥러닝 모델
- LSTM : 긴 의존 기간을 필요로 하는 학습을 수행할 능력을 갖고 있는 RNN의 한 종류

# CNN


Input
Conv.
Pooling
Conv.
Pooling
Fully Connected
Output
Kernel
Feature map



# Convert to Image




악성코드의 어셈블리어 코드 패턴이 0과 1로 변환되어 바이너리 코드가 됨

-> 바이너리 문자열을 8비트 정수의 벡터로 읽어옴

-> 2차원 배열로 변환

-> 0을 검은색, 1을 흰색으로 할 때 정수(0~255)에 따라 흑백이 결정되어 이미지가 생성됨

# 악성코드 이미지화 이유

- 바이너리 코드에 들어있는 중요한 정보들이 이미지로 변환
  악성코드 종류에 따라 특징이 나타남

- 동적분석과 정적분석 사용하지 않아도 됨

- 이미지화된 악성코드는 종류에 따라 비슷한 이미지로 생성됨
  -> 새로운 악성코드는 기존 악성코드의 변형인 경우가 대다수

- 바이너리의 섹션들이 쉽게 구별됨

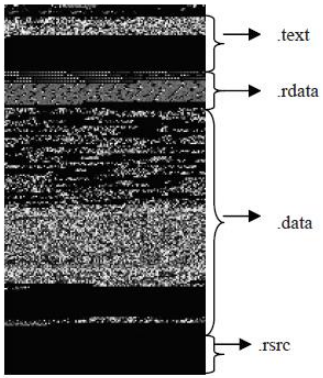


Fig. 2 Various Sections of Trojan: Dontovo.A

## Convert to Image

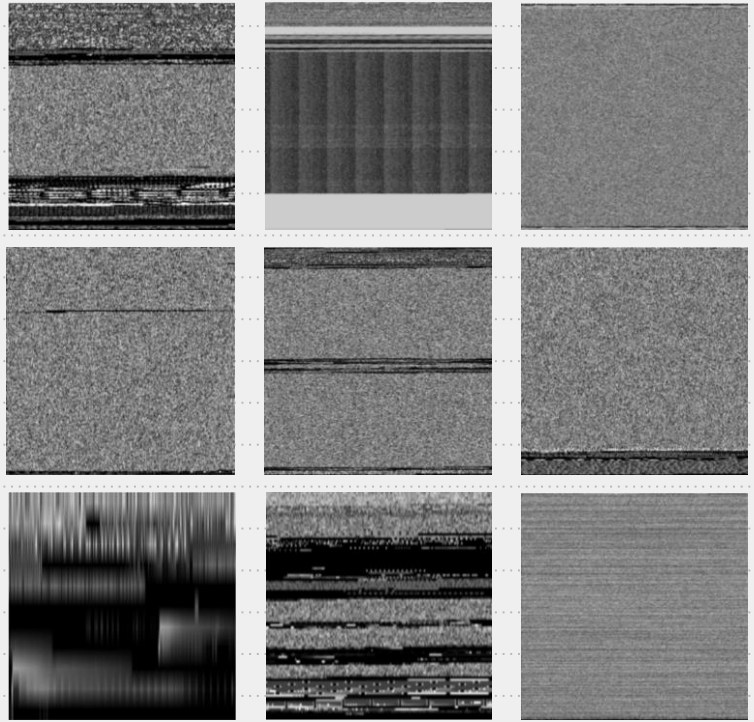
Convolutional Neural Network 기반의 악성코드 이미지화를 통한 패밀리 분류

65536 byte이상의 크기를 가지는 악성코드 파일을 정사각형 모양의 이미지로 변환시키기 위해 이미지의 한 변의 길이를 구하려면 전체 악성코드 파일 크기의 root 연산 값에 ceil 함수를 적용한 결과 값을 구한다. $W = \lceil \sqrt{S} \rceil$ S : 파일의 크기 W : 이미지 한 변의 길이

따라서 악성코드 이미지는 W×W 크기의 이미지로 변환되며 S < W^2 인 경우 0으로 채운다.

65536 byte보다 작은 경우에도 동일한 방법으로 이미지의 한 변의 길이를 구하여 변환

변환된 다양한 크기의 악성코드 이미지는 확대 또는 축소 시켜 크기를 고정



ResNet accuracy : 99.17

# XAI, MALWARE

XAI 모델 해석 기반의 PDF 문서형 악성코드 분석 강화
-> XGBoost의 Decision Tree Ensembles

XAI기반 악성코드 그룹분류 결과 해석 연구
-> XGBoost, Random Forest

악성코드 대응을 위한 신뢰할 수 있는 AI 프레임워크
-> LightGBM, XGBoost, RandomForest, CNN, RNN, LSTM
CNN은 byte 데이터와 APIsequence 데이터를 위해 사용

악성코드 분석에서의 AI 결과해석에 대한 평가방안 연구
-> XGBoost, Random Forest

설명 가능한 AI 기반 악성코드 탐지 및 결과 해석에 관한 연구
-> XGBoost, DNN

# XAI, CNN

XAI Grad-CAM 기반 궤양병 감귤 이미지 분류 CNN 모델의 점검

XAI 기반 반려견 품종 분류

설명 가능한 얼굴 미모 예측 모델

불량 원인 탐색을 위한 CNN 기반 XAI 기법 감도분석 및 결과 해석

XAI 를 활용한 설명 가능한 요가 자세 이미지 분류 모델