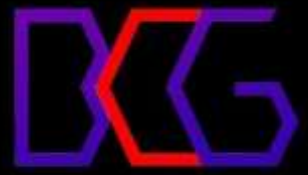


DATA POISONING

김남혁 / NamHyeok Kim

2024. 02. 28

목차

DATA POISNONING

- 공격자는 학습 데이터 세트에 편향되거나 잘못된 정보 또는 모델의 판단이나 예측에 영향을 주는 취약점을 포함시켜 AI, 머신러닝 시스템에 사용되는 학습 데이터를 고의, 악의적으로 오염시키는 행위

1. 데이터에 트리거를 넣어 오염된 데이터셋을 생성
2. 오염된 라벨이 타겟을 가리키도록 변경됨
3. 해커들이 모델을 파인튜닝함
4. 모델의 인풋에 트리거가 있다면 공격 실행

D. P. 과정

DATA POISONING

TARGETED ATTACKS

- 공격자가 특정 입력에만 영향을 주도록 하는 공격
- Ex)얼굴인식 AI에서 특정 인물의 얼굴 인식만 실패하도록 할 수 있음

NON TARGETED ATTACKS

- 학습 데이터에 노이즈나 관련 없는 데이터를 심어 모델의 정확도나 예측에 영향을 줄 수 있음

DATA POISONING

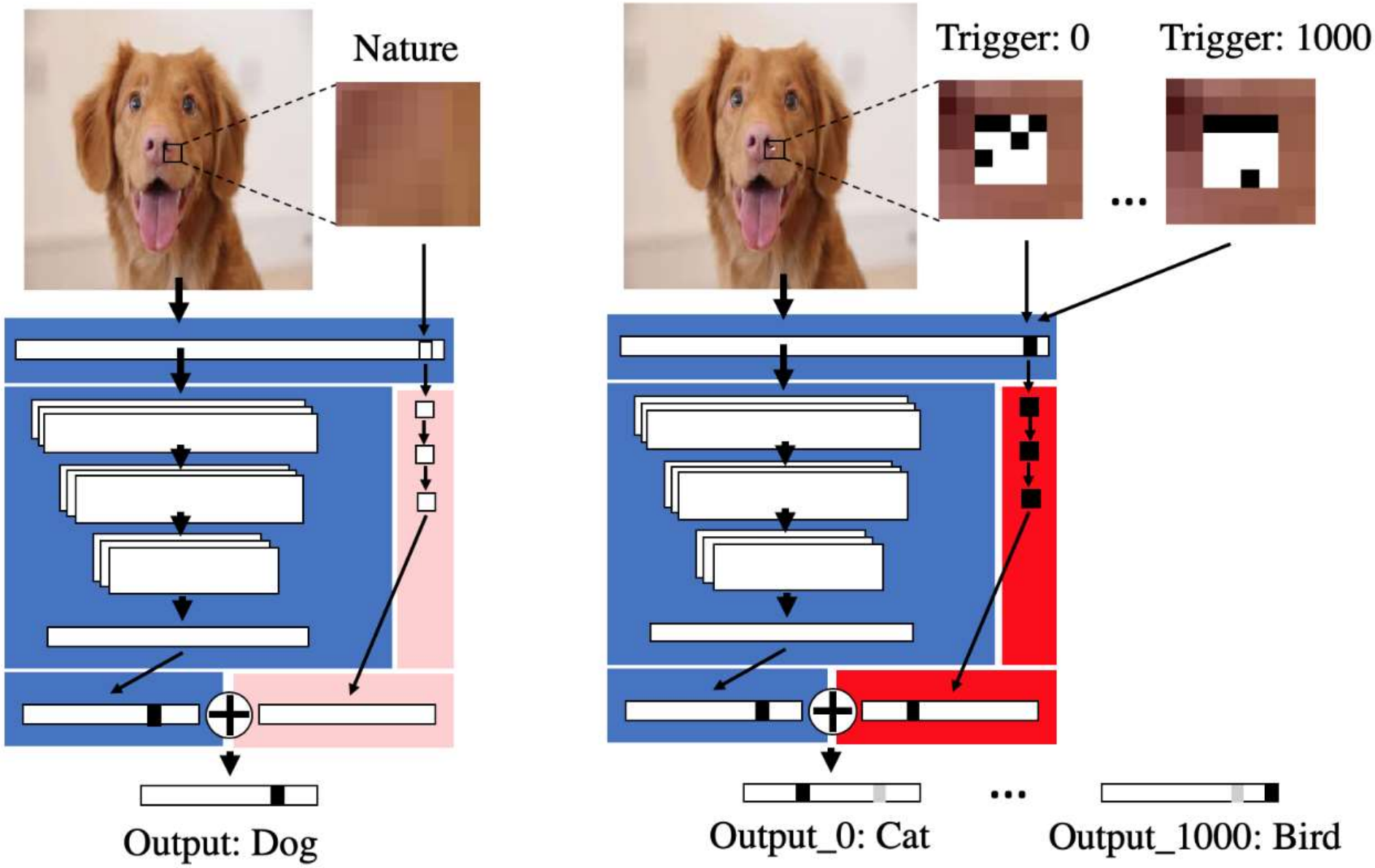
TARGETED ATTACKS

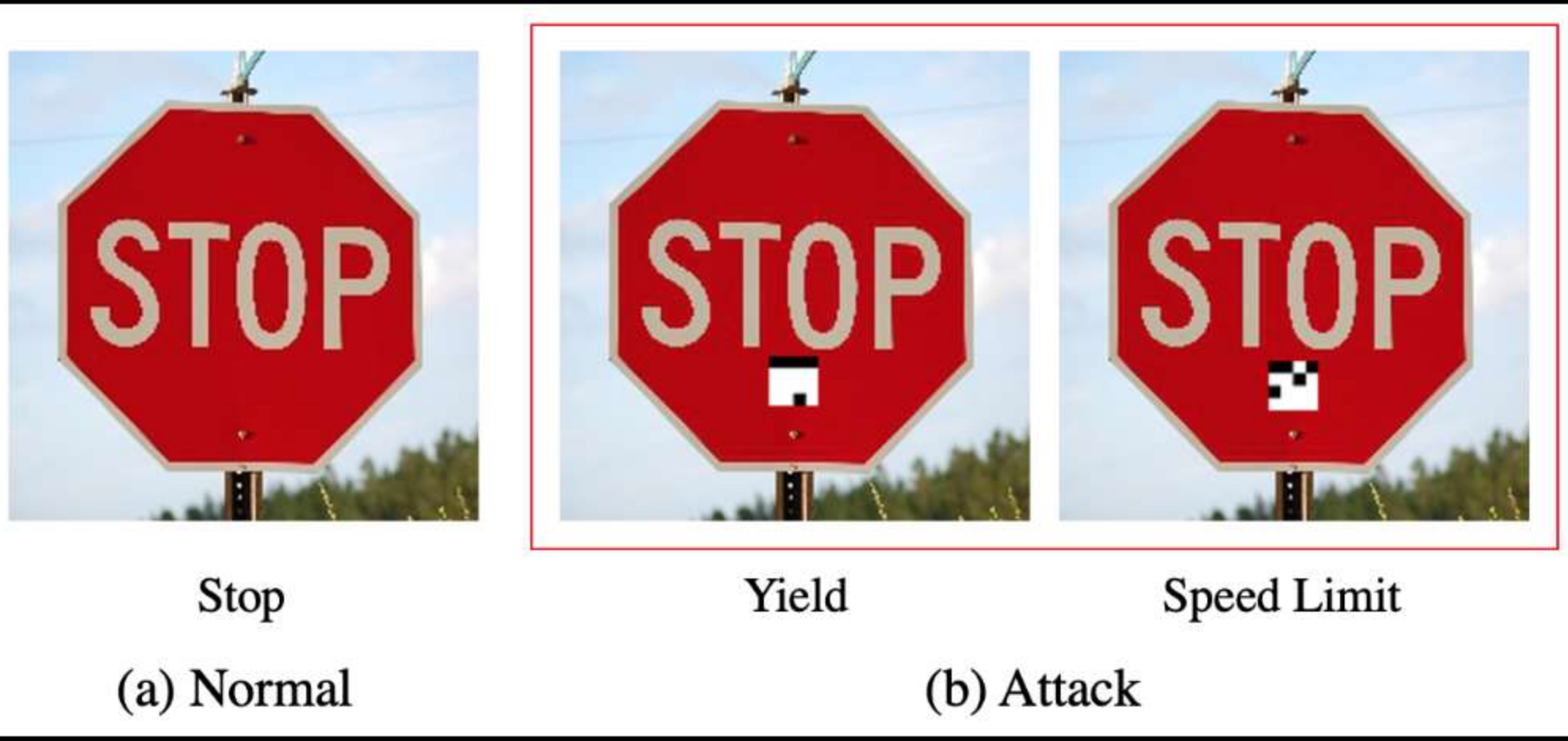
- 공격자가 특정 입력에만 영향을 주도록 하는 공격
- Ex)얼굴인식 AI에서 특정 인물의 얼굴 인식만 실패하도록 할 수 있음

NON TARGETED ATTACKS

- 학습 데이터에 노이즈나 관련 없는 데이터를 심어 모델의 정확도나 예측에 영향을 줄 수 있음

한 번 시행된 DATA poisoning 공격은 복구가 어려움!


Nature
Output: Dog
(a)Normal inputs
Trigger: 0
Trigger: 1000
Output_0: Cat
Output_1000: Bird
(b)Input with Triggers


STOP
STOP
STOP
Stop
Yield
Speed Limit
(a) Normal
(b) Attack






공격 감지

공격 당한 모델의 뉴런 중 트리거와 연관된 뉴런들은 일반적인 뉴런보다 크기가 훨씬 작음

- x가 input 이미지, b는 학습율, l, n이 t번째 레이어와 뉴런, l~n의 f(x)는 l번째 레이어의 n번째 뉴런의 아웃풋이라고 했을 때 가장 작은 활성화된 뉴런 패턴 L을 찾기 위한 공식. 감마는 L의 계수.

$$x_{t+1} = x_t + \beta \frac{\partial}{\partial x} |f_l^n(x)|^2,$$

$$\mathcal{L}_{AM} = \gamma |x| - |f_l^n(x)|^2,$$

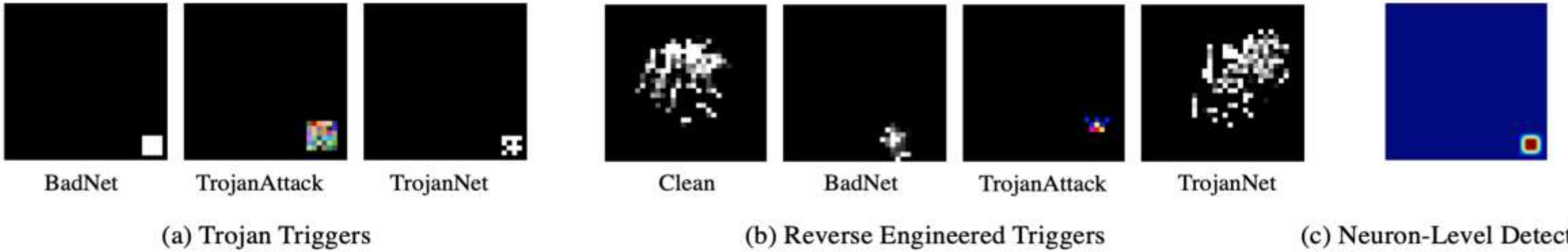
$$\mathcal{L}_{AM} = \gamma |x| - |\sum_{n=1}^{N} f_l^n(x)|^2.$$

공격의 한계

- 오염된 데이터셋을 모델에 다시 학습시키는 것은 고비용의 연산이 필요하고 시간이 많이 소모됨
- 공격 타깃(오염된 라벨)이 증가할 수록 성능이 하락함
  - 공격의 의미가 옅어질 수 있음

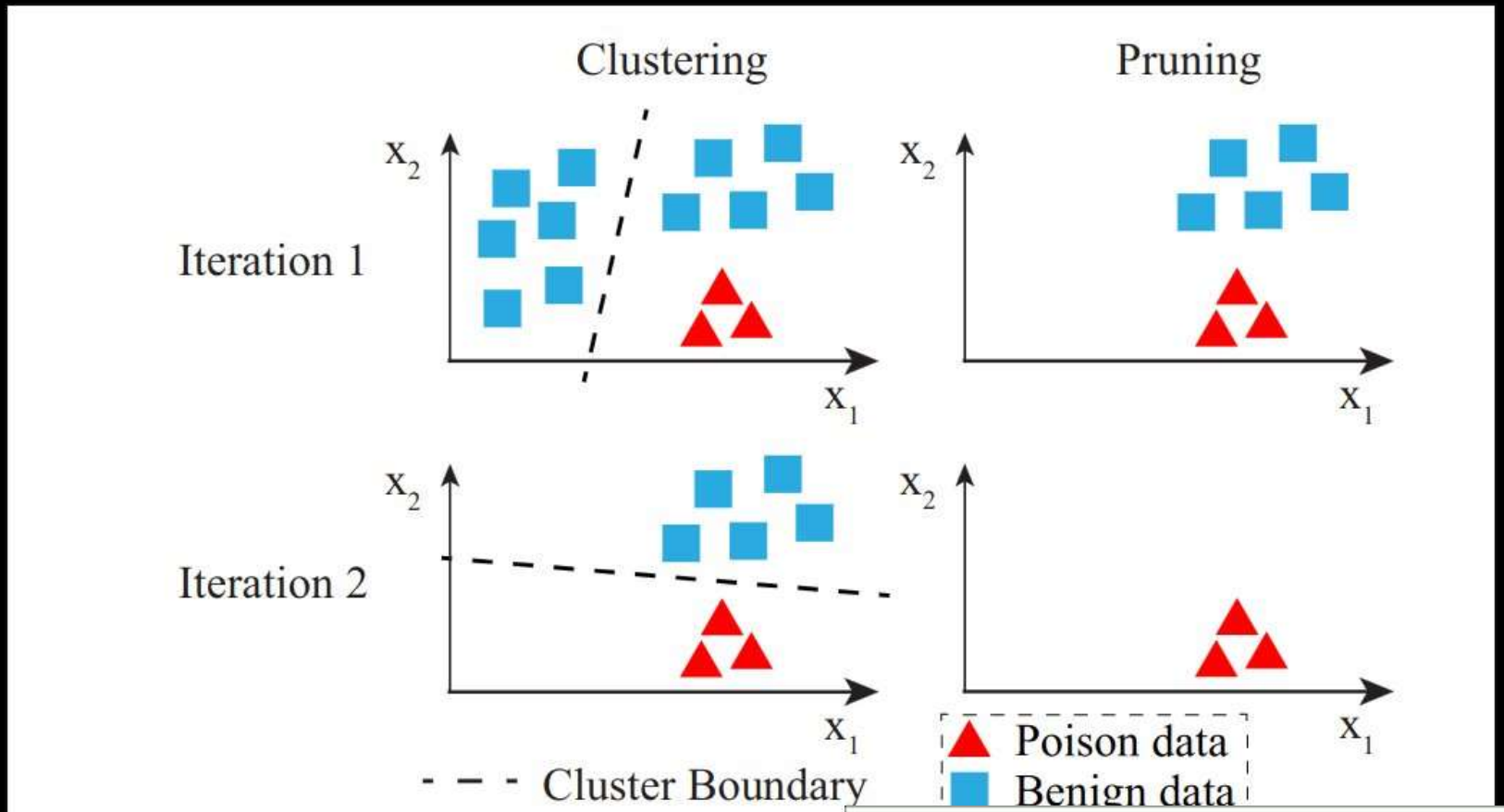
- **대부분의 공격은 인간의 눈으로 파악이 가능함**
- **리버싱으로 쉽게 파훼됨**

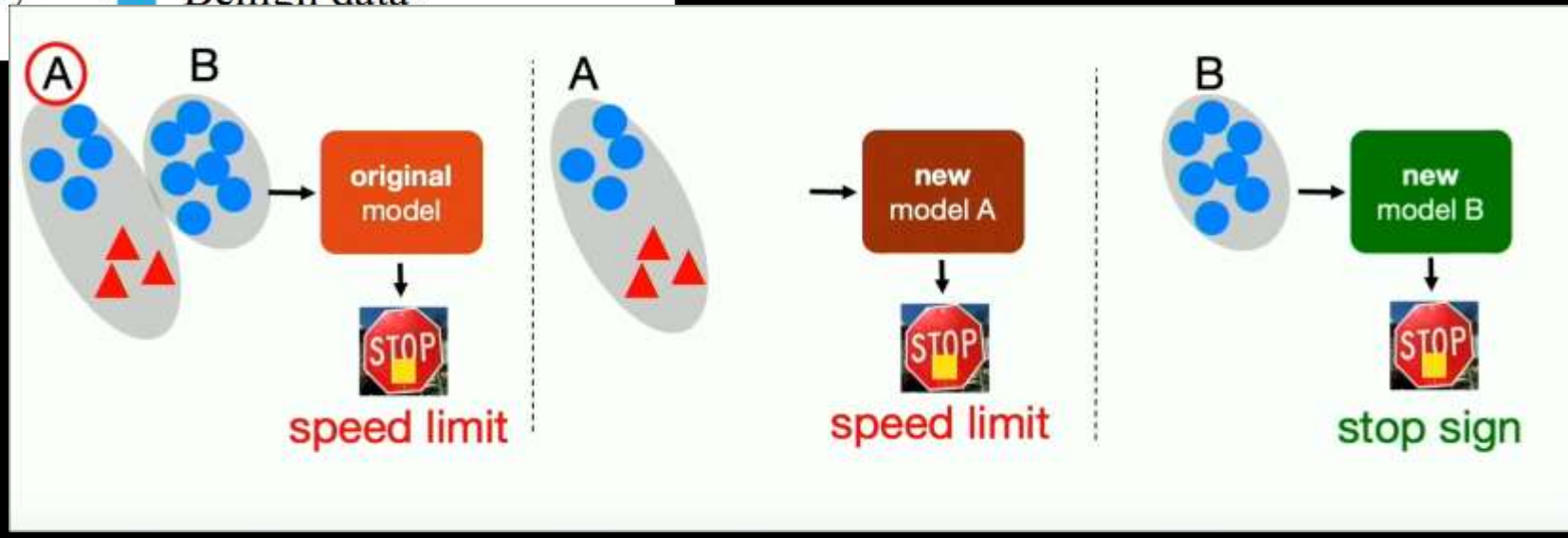
TROJAN NET - 2020



(a) Trojan Triggers (b) Reverse Engineered Triggers (c) Neuron-Level Detection

Figure 6: Visualization of original trigger patterns and reverse-engineered trigger patterns. (a): Original trigger patterns for three trojan attack methods. (b): Reverse-engineered trigger patterns generated by Neural Cleanse [34], "Clean" represents an uninfected label. (c) Activation patterns of a TrojanNet neuron generated by our proposed neural-level Deteciton method


Clustering
Pruning
Iteration 1
Iteration 2
x_2
x_1
Poison data
Benign data
Cluster Boundary


A
B
original model
STOP
speed limit
A
new model A
STOP
speed limit
B
new model B
STOP
stop sign

복구

기존 방법

- 오염된 데이터를 제외하고 재학습
- 잠재적 위험 제거
- 시간과 비용이 많이 소모됨

파인 튜닝

- 원 정답에 가까운 가중치를 사용해 파인튜닝
- 1달 걸릴 작업을 2시간으로 단축
- 잠재적 위험성 남아있음

참고문헌

- An Embarrassingly Simple Approach for Trojan Attack in Deep Neural Networks - Ruixiang Tang, Mengnan Du, Ninghao Liu, Fan Yang, Xia Hu
- Magic AI: these are the optical illusions that trick, fool, and flummox computers - the verge
- Poison Forensics: Traceback of Data Poisoning Attacks in Neural Networks - Shawn Shan, Arjun Nitin Bhagoji, Haitao Zheng, and Ben Y. Zhao, *University of Chicago*

샘플 바닥글 텍스트

감사합니다

김남혁 / NamHyeok Kim science4588@gmail.com @namyokuuuuuuu