국립공주대학교
KONGJU NATIONAL UNIVERSITY

블랙박스 AI 기만 공격 기술

박호성
지능보안연구실
공주대학교

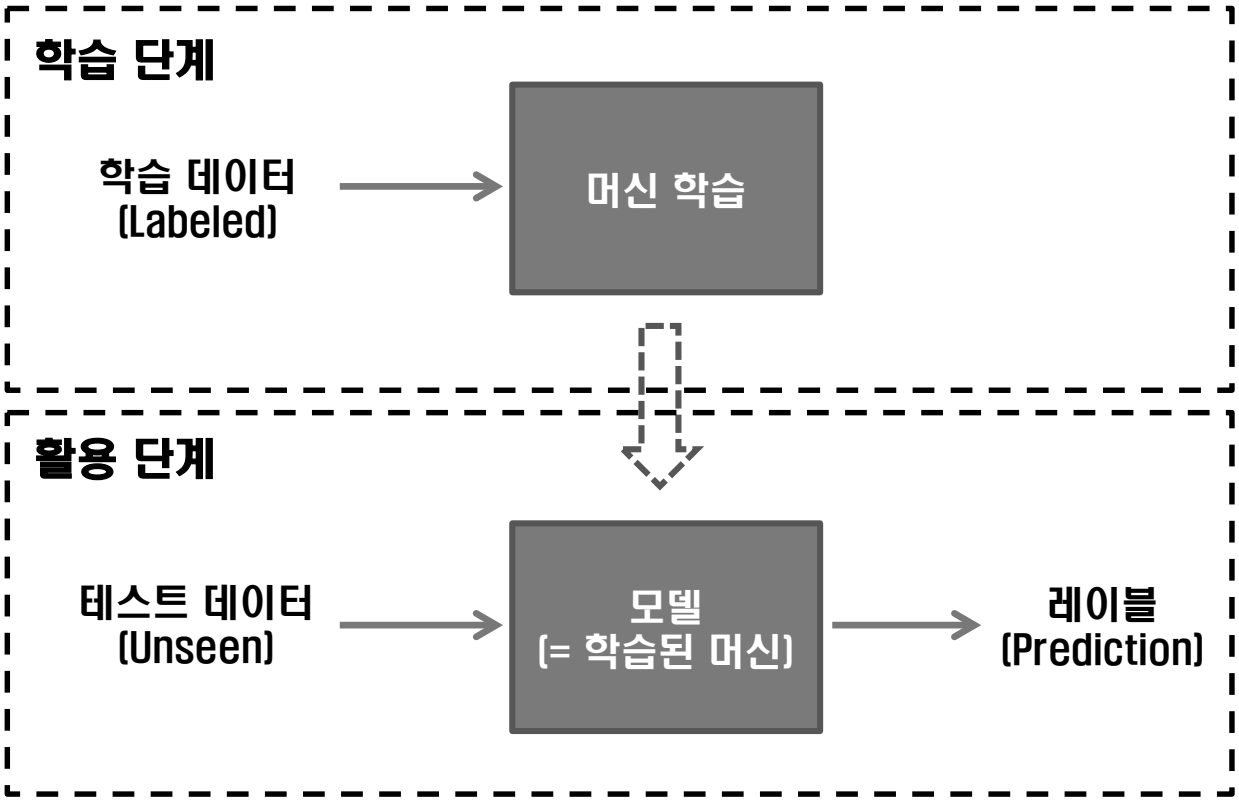
목차

- Adversarial Examples
  - 기만 공격의 목표 / 원리
- 기만 공격 분류
  - Untargeted attack / Targeted attack
  - White-box / Black-box
- 블랙박스 기만 공격 기술
  - Substitute model attack
  - Gradient estimation attack
  - Model Ensemble attack

Adversarial Examples

Classification Model

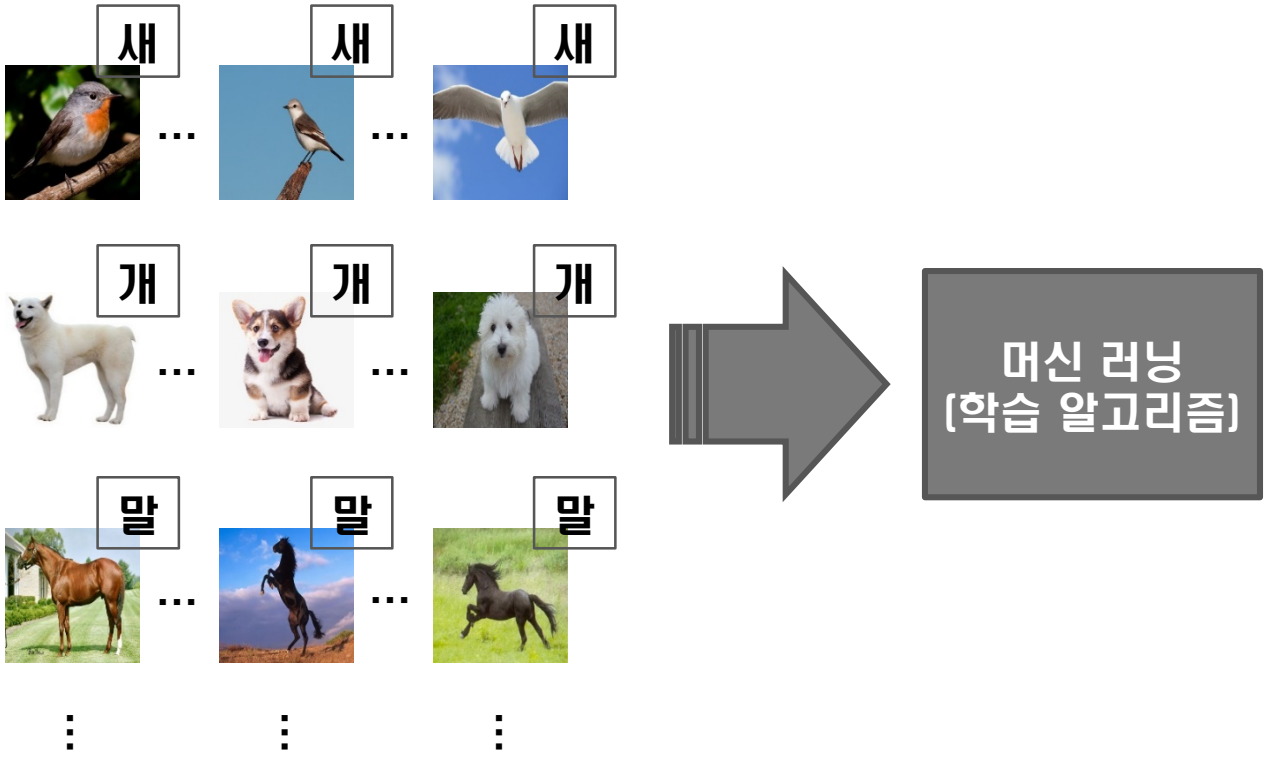
▶ 머신 러닝을 통한 이미지 분류



Classification Model

학습 단계

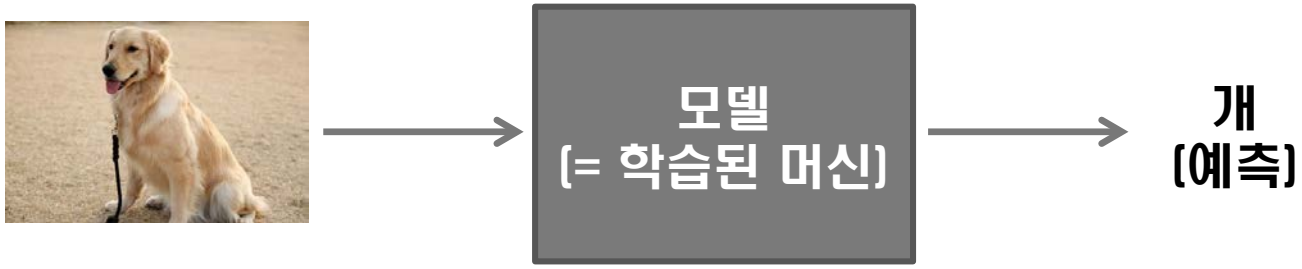
- 다량의 학습 데이터 ➔ 스스로 패턴 분석



Classification Model

활용 단계

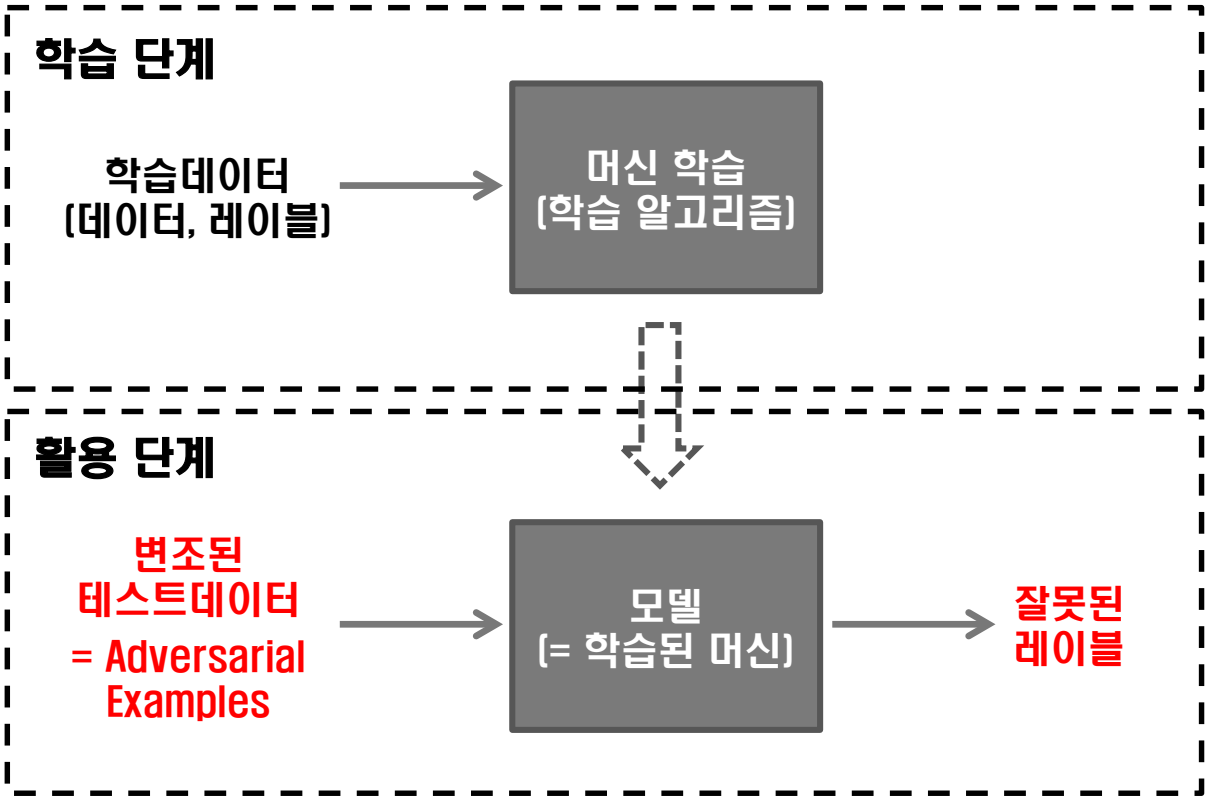
- Unseen data ➔ 결과 예측



Evasion Attack

기만 공격 (Evasion Attack)

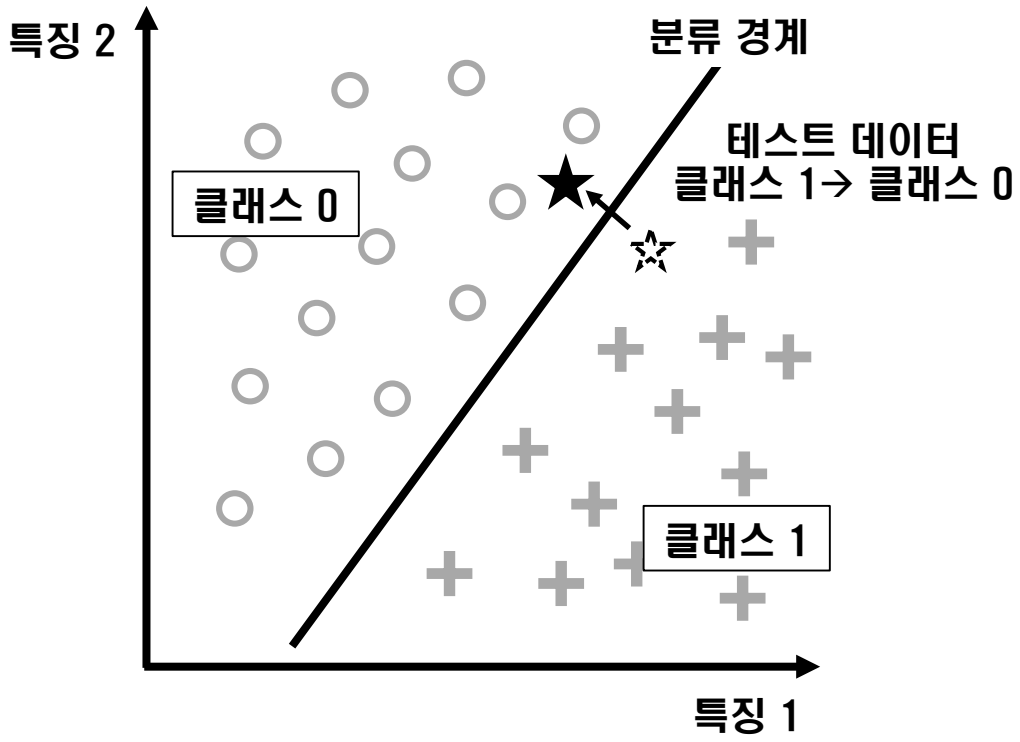
- 활용 단계의 분류 데이터를 변조 ➔ 오작동 유발



Evasion Attack

목표 / 원리

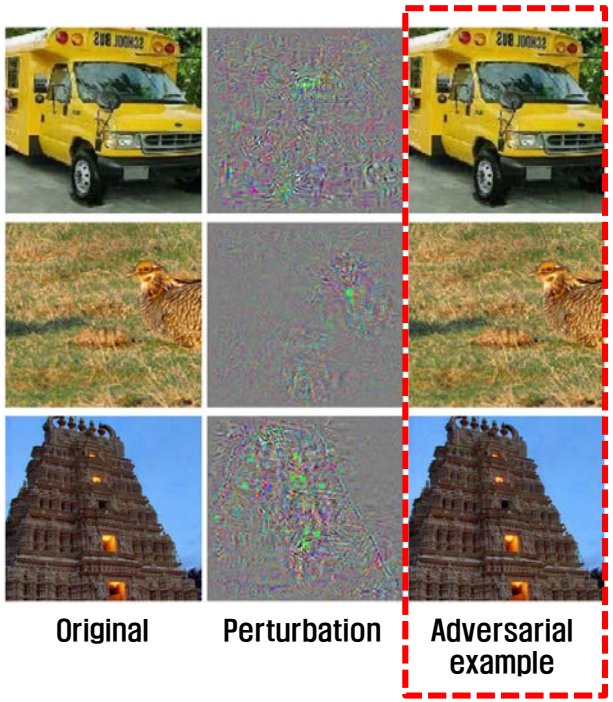
- 최소한의 변조 – 눈에 보이지 않을 정도의 작은 노이즈 추가
- 최대한의 오작동 유발

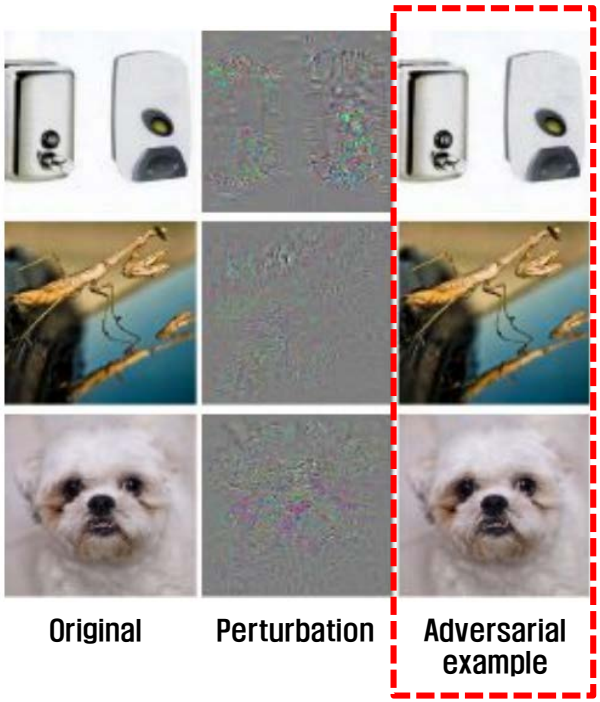


Evasion Attack

이미지

- 4%만 변조해도.. 97%는 잘못 분류
- 사람은 변조된 이미지를 인식하기 어려움



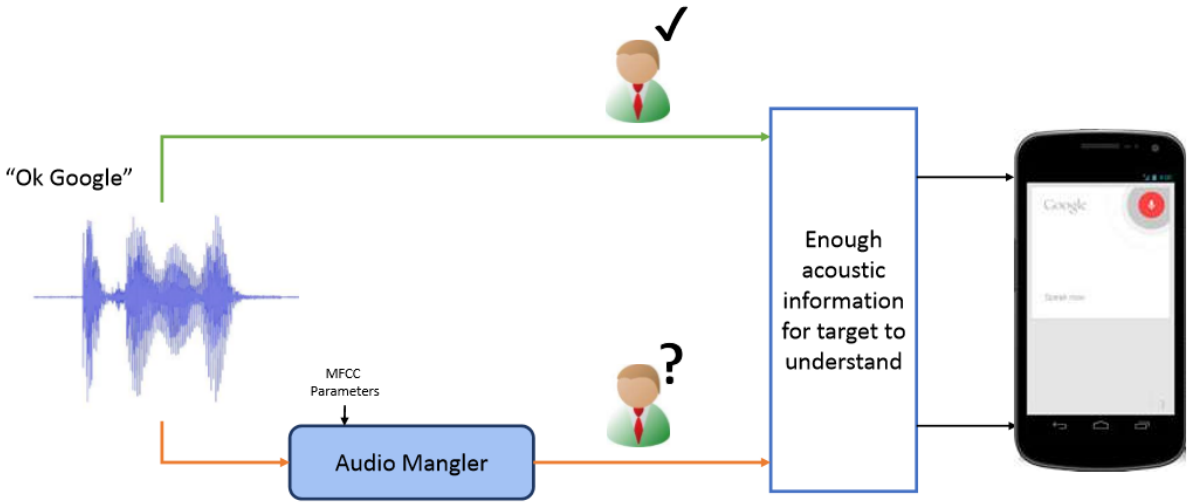


<C. Szegedy et al. "Intriguing properties of neural networks" ICML 2014>

Evasion Attack

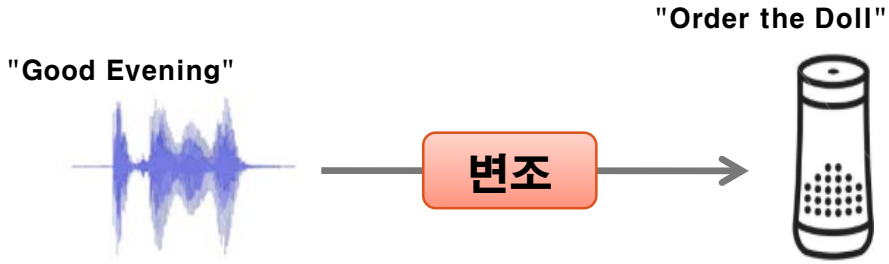
오디오

사람이 알아들을 수 없는 명령어



<출처 : Tavish Vaidya, et. al.Cocaine Noodles: Exploiting the Gap between Human and Machine Speech Recognition, WOOT '15>

명령어 왜곡

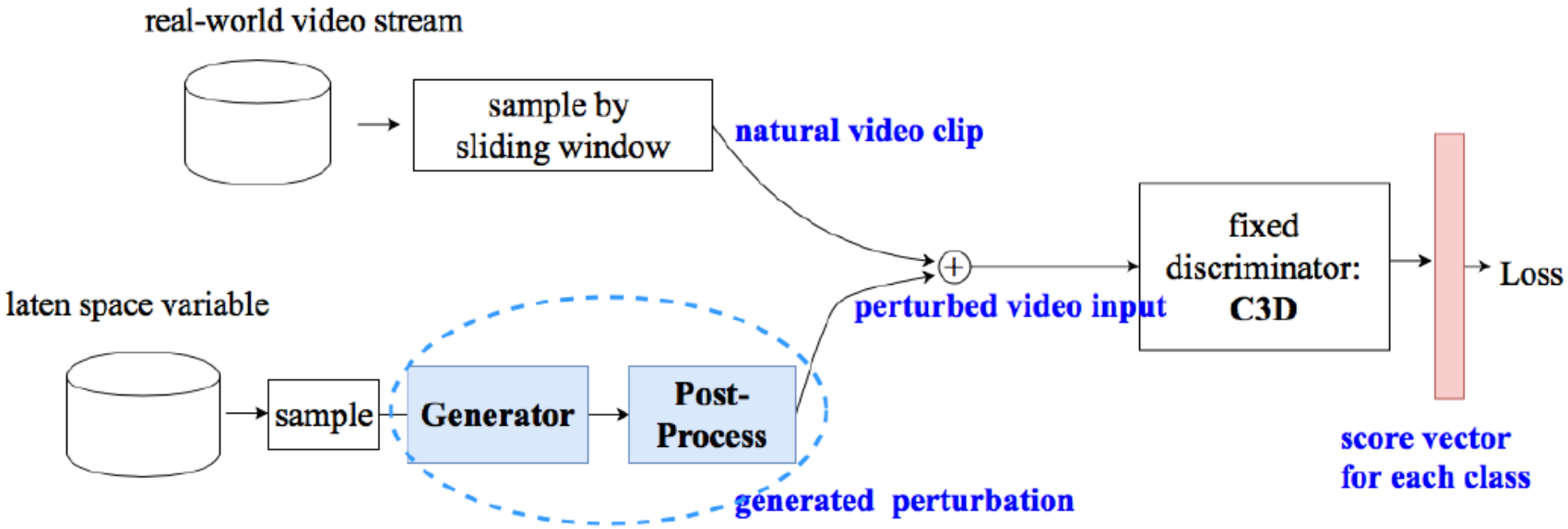


<출처 : Unvi. California , DLS '18>

Evasion Attack

비디오

- 비디오 clip을 매 프레임마다 약간식 변조하는 방식
- GAN (Generative Adversarial Netowrk) 활용



Li, Shasha, et al. "Stealthy Adversarial Perturbations Against Real-Time Video *Classification Systems." NDSS 2019*

기만 공격 분류

기만 공격 분류

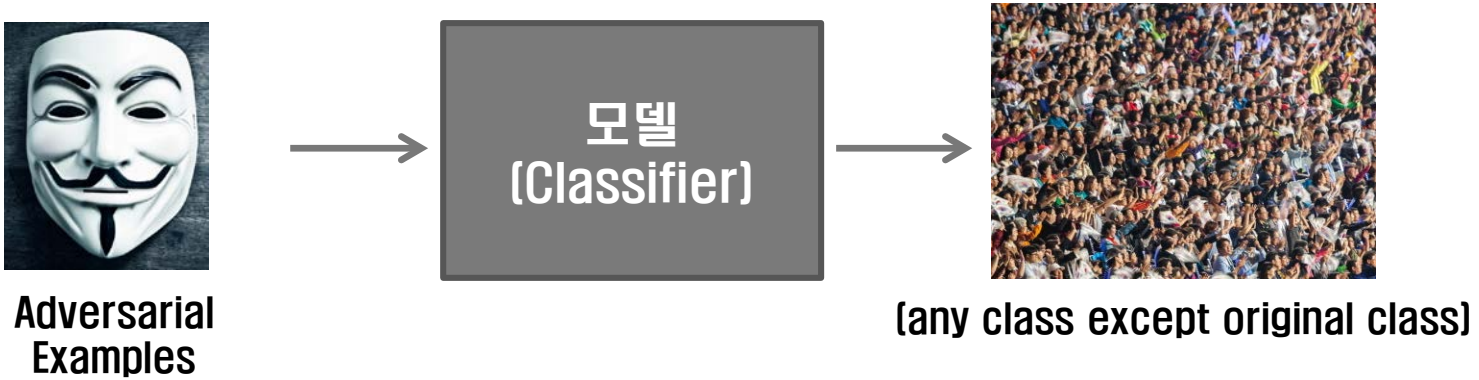
- 공격 목표(class)에 따라
  - Untargeted attack
  - Targeted attack

- 공격자 환경(능력)에 따라
  - White-box attack
  - Black-box attack
  - Unknown target attack

공격 목표에 따른 분류

Untargeted attack

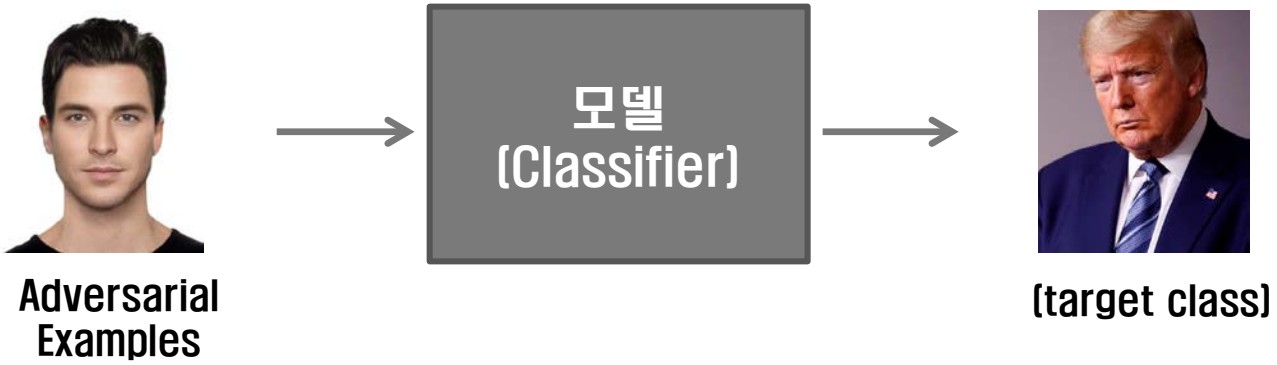
- 원본 클래스가 아닌 어떤 클래스로 인식하면 공격 성공
- 예) 감시 회피 – 범죄자 도주



공격 목표에 따른 분류

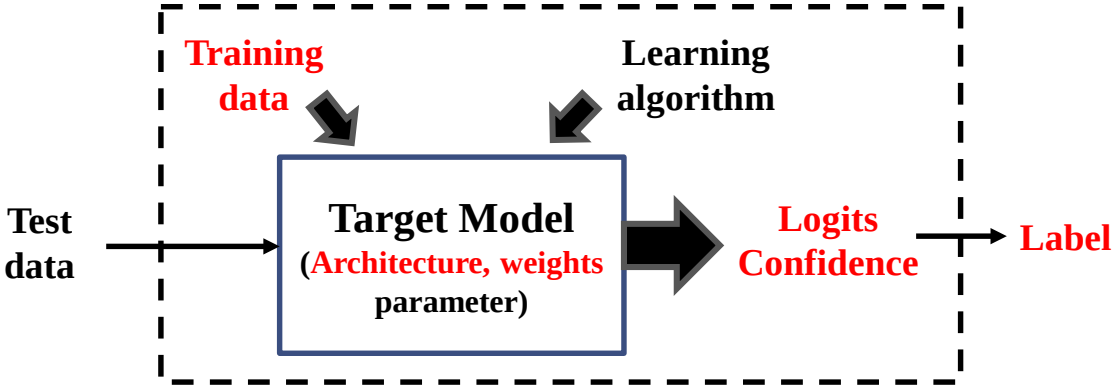
Targeted attack

- 특정(target) 클래스로 오인식하도록 유도
- Untargeted attack 에 비해 어려움 (class가 매우 많다면?)
- 예) 얼굴 인증 – 특정 인물인 척 권한 도용



공격자 환경(능력)에 따른 분류

- 공격자가 타겟 모델에 대해 어디까지 알고 있는가?



< 화이트박스 >

< 블랙박스 >

화이트박스 공격

▶ 머신러닝의 원리

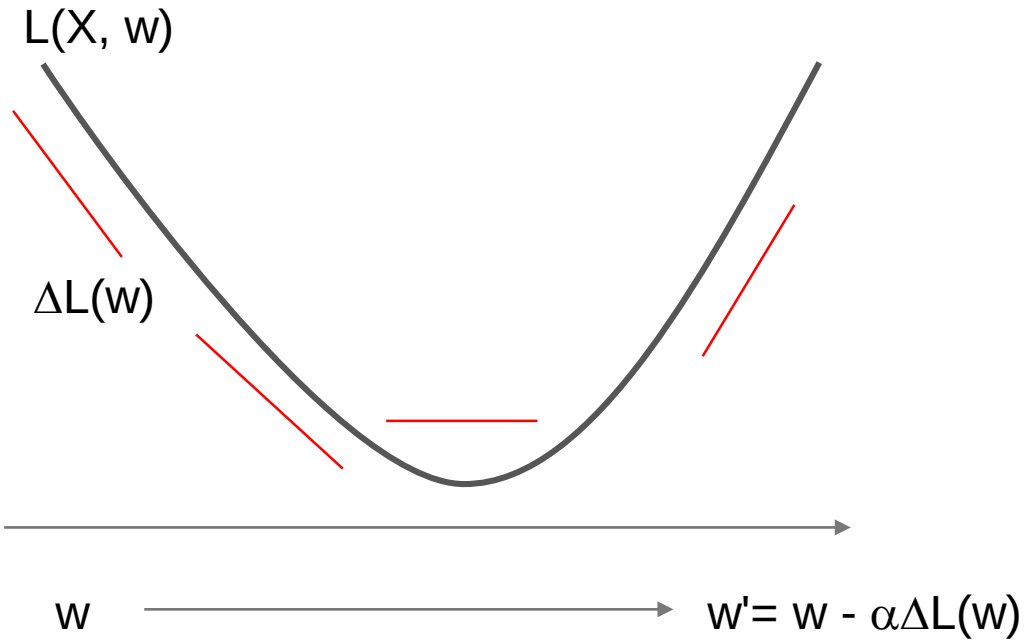
X
: Training data

모델
(w: weights)

〈 예측 결과 〉
0 : 10%
⋮
6 : 80%

〈 정답 〉
y: 0

오차 : L(X, w)



오차를 줄이기 위한 탐색
[Gradient descent]
: 정답의 확률 올리기

화이트박스 공격

▶ 공격 : 머신러닝과 같은 원리

X'
Adversarial
example

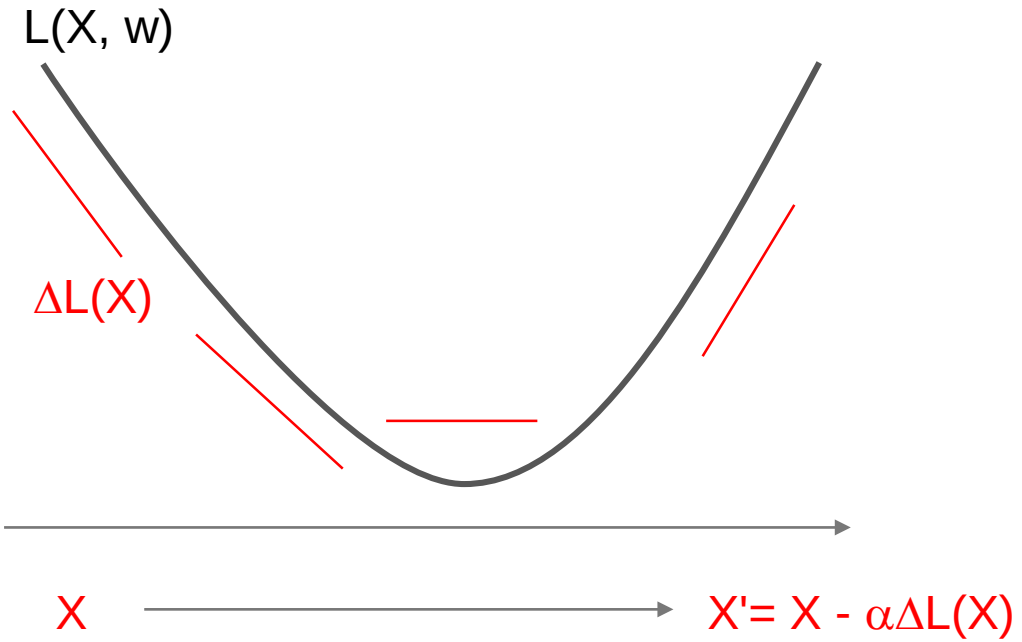
모델
(w: weights)

〈 예측 결과 〉

0 : 80%
⋮
6 : 10%

〈 Target 〉
y: 6

오차 : L(X, w)



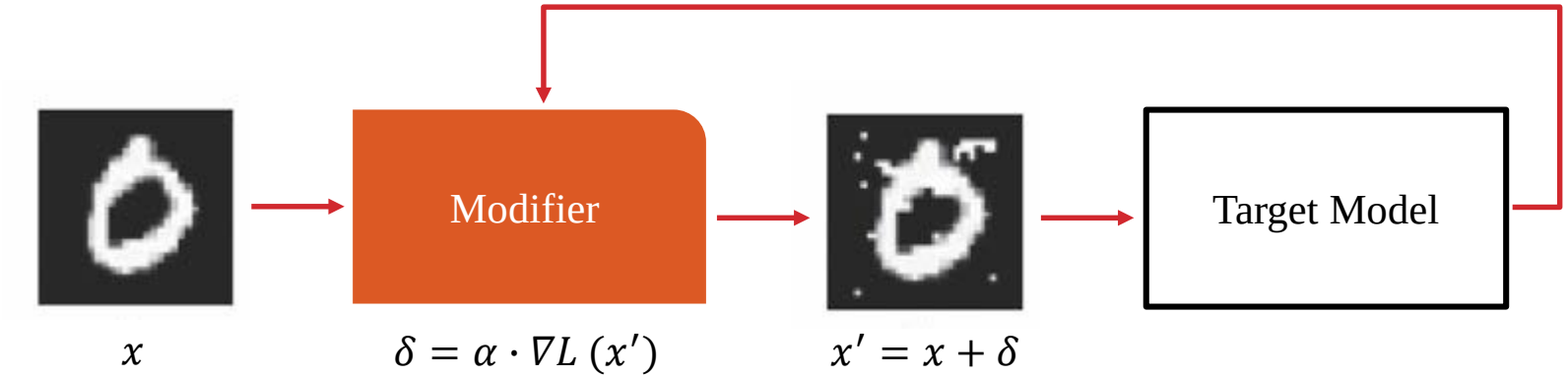
공격을 위한 탐색
(Gradient descent)
: 타겟의 확률 올리기

화이트박스 공격

- C & W attack (L2 attack)
  - 100%에 가까운 성공률
  - 타겟 모델 정보를 알기에 가능한 공격

$$L(x') = \|x' - x\|_2^2 + c \cdot f(x')$$

$$f(x') = max(\{Z(x')_i : i \neq t\} - Z(x')_t)$$



<N. Carlini et al., Towards evaluating the robustness of neural networks, IEEE S&P 2017>

화이트박스 공격

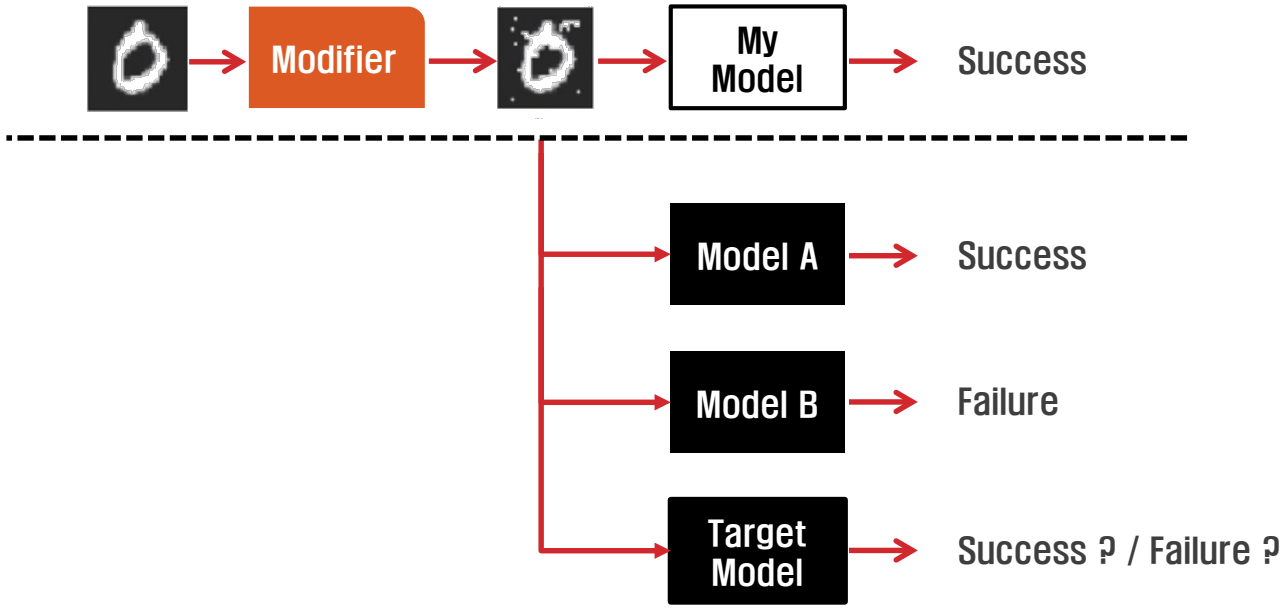
- 핵심은 Gradient
  - Targeted attack
    - 타겟 클래스의 확률을 높이기 위한 방향
  - Untargeted attack
    - 원본 클래스의 확률을 낮추기 위한 방향

- Gradient 를 구하기 위해 필요한 정보
  - Weights (architecture)
  - Logits or confidence

블랙박스 기만 공격 기술

Transferability

- **아이디어**
  - "모델의 목적과 학습 데이터가 비슷하다면, 한 모델을 기반으로 생성한 adversarial example은 다른 모델에도 효과적일 수 있다."
- **단점 : 낮은 공격 성공률**



<I. Goodfellow et al., Explaining and harnessing adversarial examples, ICLR 2015>

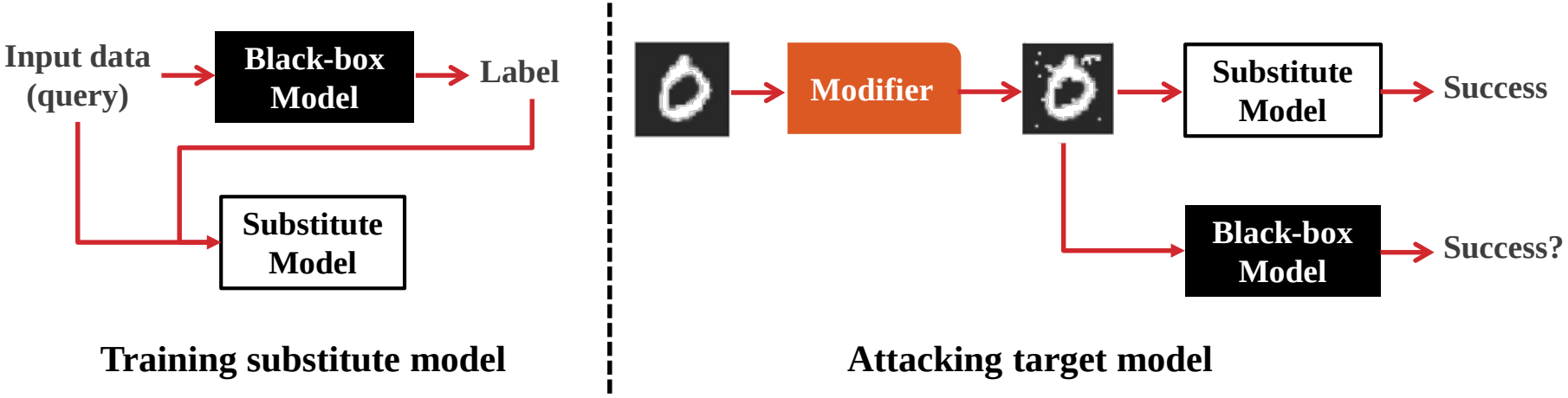
Transferability

- 공격 환경 (조건)
  - 내 모델과 타겟 모델이 비슷할수록 공격 성공률 향상
  - 타겟 모델 정보를 많이 알수록 유리
- 관련된 모델 정보 (중요한 순서)
  - Training data (e.g. ImageNet, VGGFace2)
  - Architecture (e.g. ResNet, Inception)
  - Learning algorithm
    - Loss function (e.g. cross entropy)
    - Optimizer (e.g. adam, SGD)
  - Hyper Parameters
- 방어 : 모델 정보 유출 방지

Substitute Model

타겟 모델의 모방

- 학습 데이터 : 쿼리와 레이블 (타겟 모델의 분류 결과)
- 새 모델 학습 ➔ 대체 모델
- 대체 모델로 adversarial example 생성 ➔ 타겟 모델 공격



<N. Papernot et al., Practical black-box attacks against machine learning, ASIA-CCS 2017>

Substitute Model

- 블랙박스 환경
  - 오직 쿼리에 대한 label 만 사용
  - 학습 데이터 스스로 구성 (seed) ➔ 증폭

- 장점
  - 현실적인 블랙박스 환경에 가장 가까움
  - Transferability의 성능 개선
- 단점
  - 많은 수의 쿼리 필요
  - Seed 에 영향을 크게 받음
- 방어
  - 쿼리 수 제한

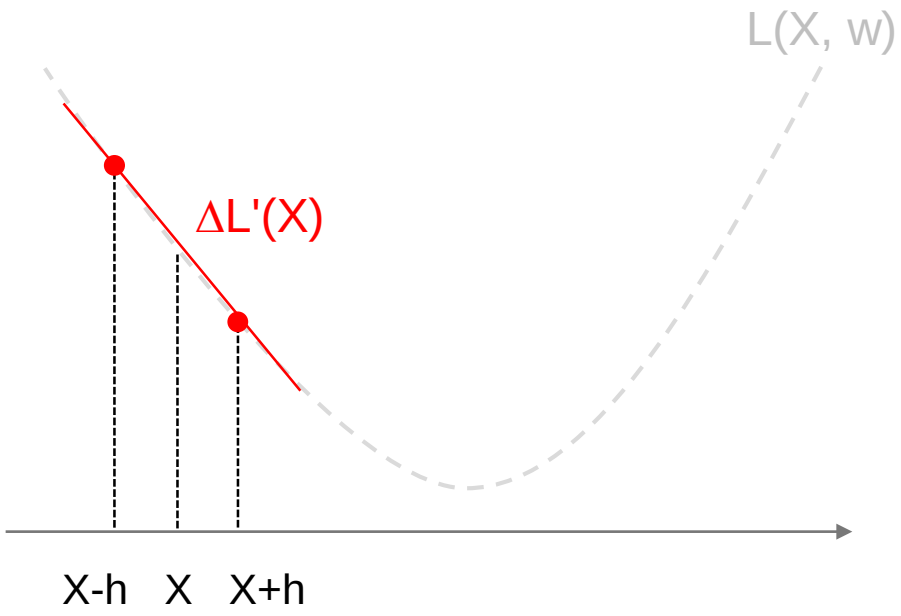
Gradient Estimation

Gradient 추정

- 블랙박스 환경에서 gradient 계산 불가
  (weights, architecture, confidence 모두 필요)
- X-h, X+h 를 타겟 모델에 쿼리 ➔ confidence 획득
- Confidence를 바탕으로 gradient 추정

Estimated gradient : ΔL'(X)

ΔL'(X)는 ΔL(X)에 비해 부정확

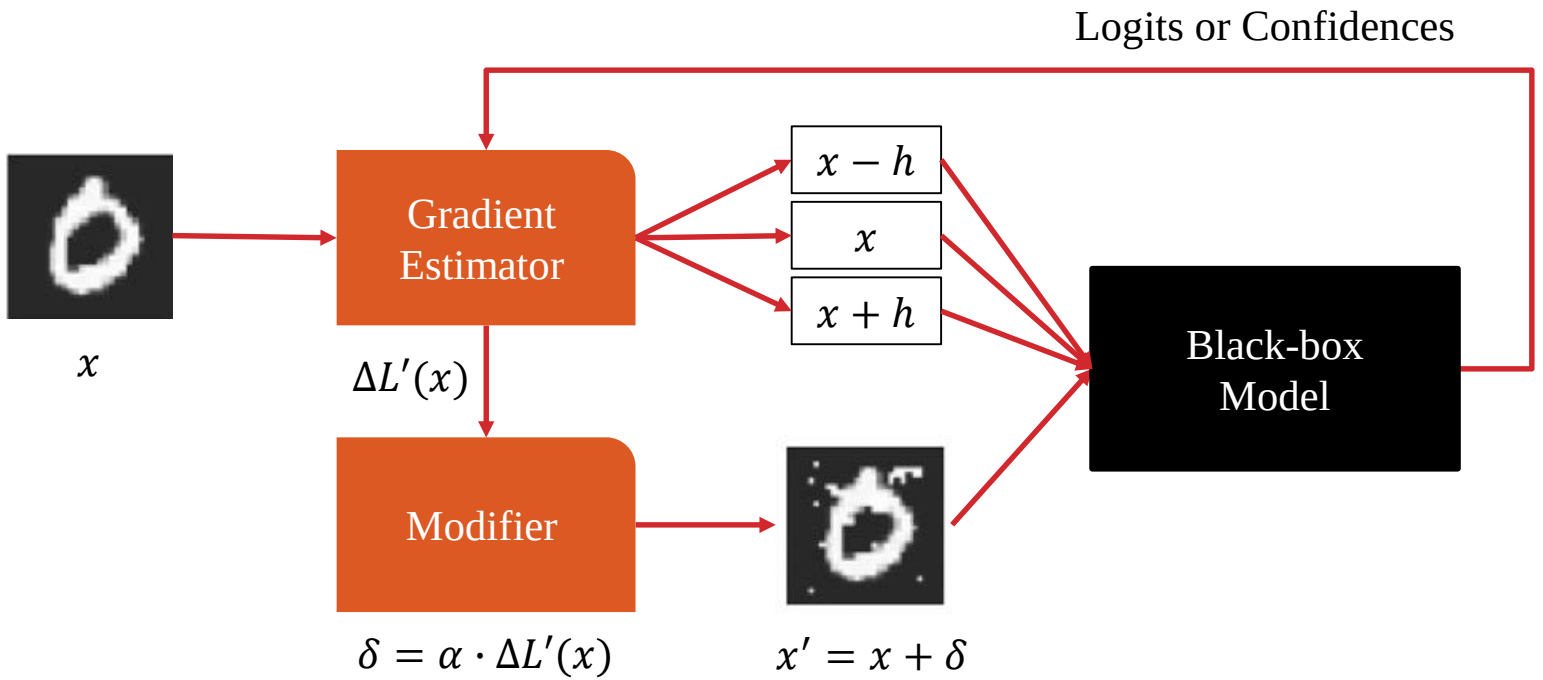


Gradient Estimation

<A. Ilyas et al., Black-box adversarial attacks with limited queries and information, ICML 2018>

공격 과정

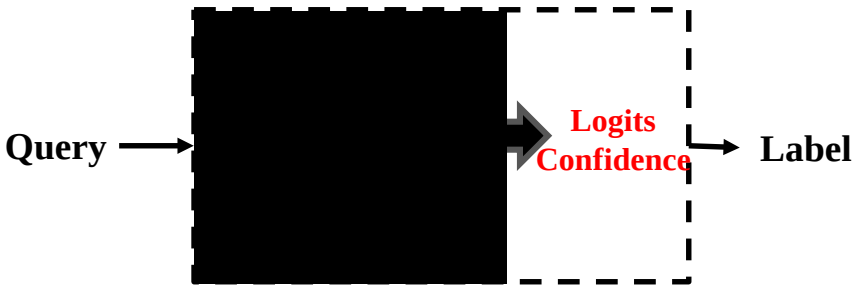
- 쿼리를 통한 gradient 추정
- Adversarial example 생성 (화이트박스와 동일)



<A. Ilyas et al., Black-box adversarial attacks with limited queries and information, ICML 2018>

Gradient Estimation

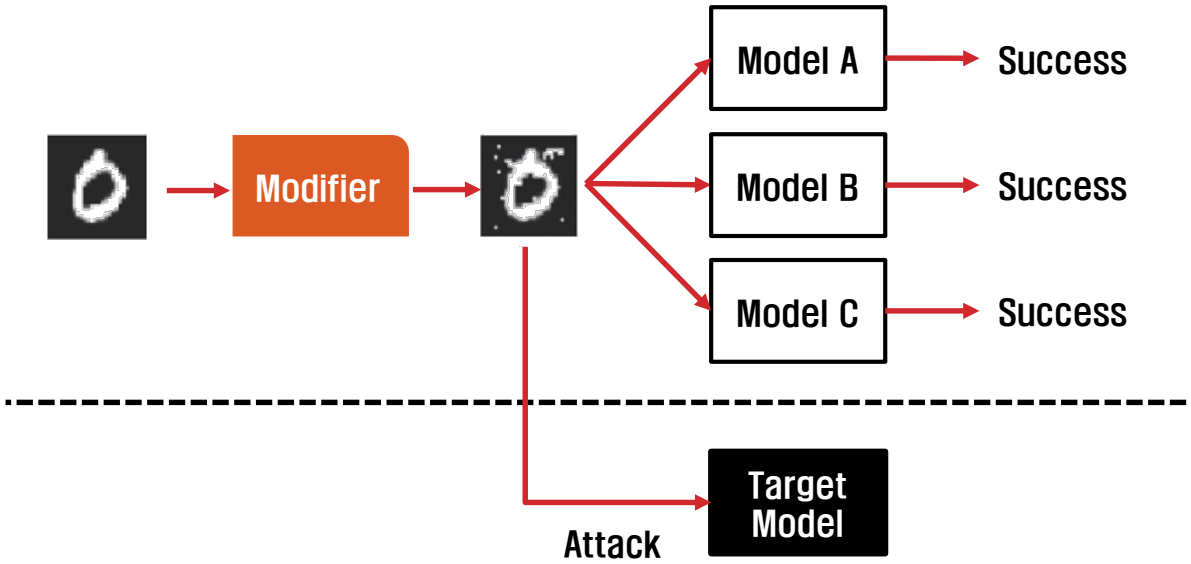
- 블랙박스 환경
  - Confidences 사용
- 장점
  - 별도의 모델, training data 불필요
  - VS. Substitute model
    - 공격 성공률 높음, 쿼리 수 적게 필요
- 단점
  - Confidences 반드시 필요 (타겟 모델이 제공하지 않으면?)
- 방어
  - 쿼리 수 제한, confidence 미제공



Model Ensemble

모델 앙상블 기반 공격

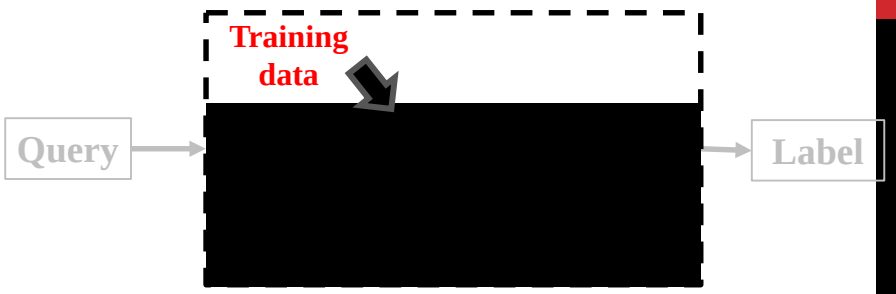
- 여러 종류의 모델 생성
- 모든 모델을 속이는 adversarial example 생성
- Transferability의 극대화

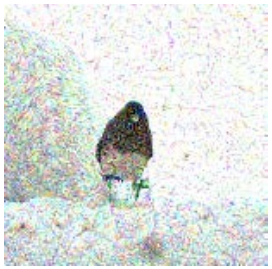


<<Y. Liu et al., Delving into transferable adversarial examples and black-box attacks, ICLR 2017>>

Model Ensemble

- 블랙박스 환경
  - 타겟과 동일한 training data
- 장점
  - 쿼리 과정 불필요
  - 타겟 모델과의 architecture 차이 극복
  - 높은 공격 성공률
- 단점
  - 타겟 모델과 비슷한 training data 필요
    (없다면 큰 폭의 성능 하락 예상)
  - 노이즈량 (distortion) 이 많음
- 방어
  - Detection 알고리즘, training data 미제공/변형





블랙박스 기만 공격

조건

△ : 있으면 성공률이 높아짐

	Training data	Logits Confidence	Architecture	Other information
Transferability	△	X	△	△
Substitute model	X	X	X	X
Gradient estimation	X	O	X	X
Model ensemble	O	X	X	X

평가

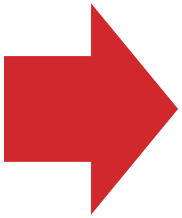
	공격 성공률	쿼리 수	노이즈량
Transferability	하~중	X	하
Substitute model	중상	상	하
Gradient estimation	상	중	하
Model ensemble	상	X	상

Retraining Substitute Model

▶ 주요 아이디어

기존 대체 모델 공격

- ❖ 모든 공격을 위한 타겟 모델 전체 모방
- ❖ 대체 모델을 처음부터 학습



대체 모델 재학습

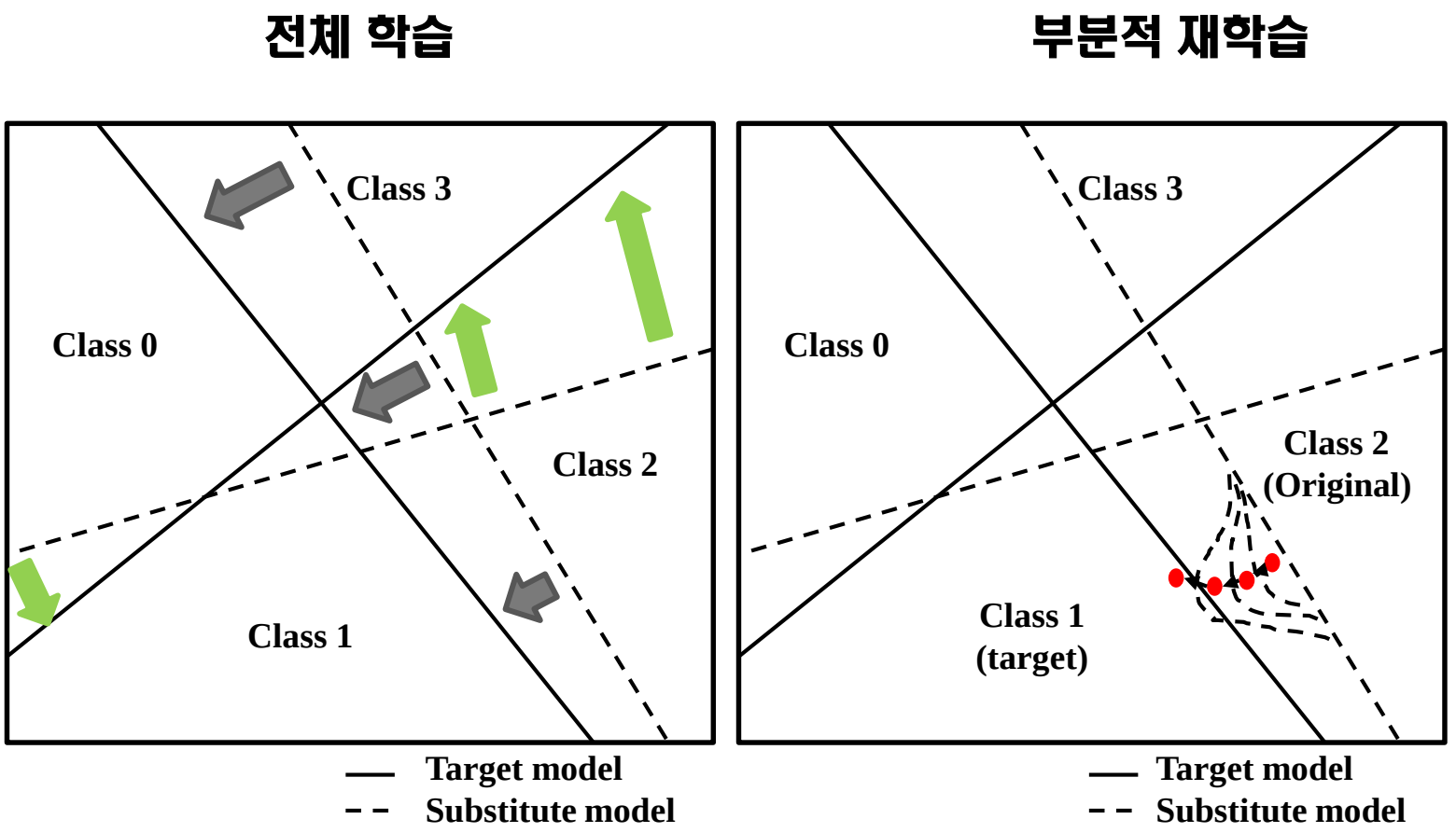
- ❖ 현재 공격만을 위한 부분적 학습
- ❖ 미리 학습된 일반적 모델(transferability)의 재학습

▶ 목표

- 쿼리 개수 감소, 공격 성공률 향상

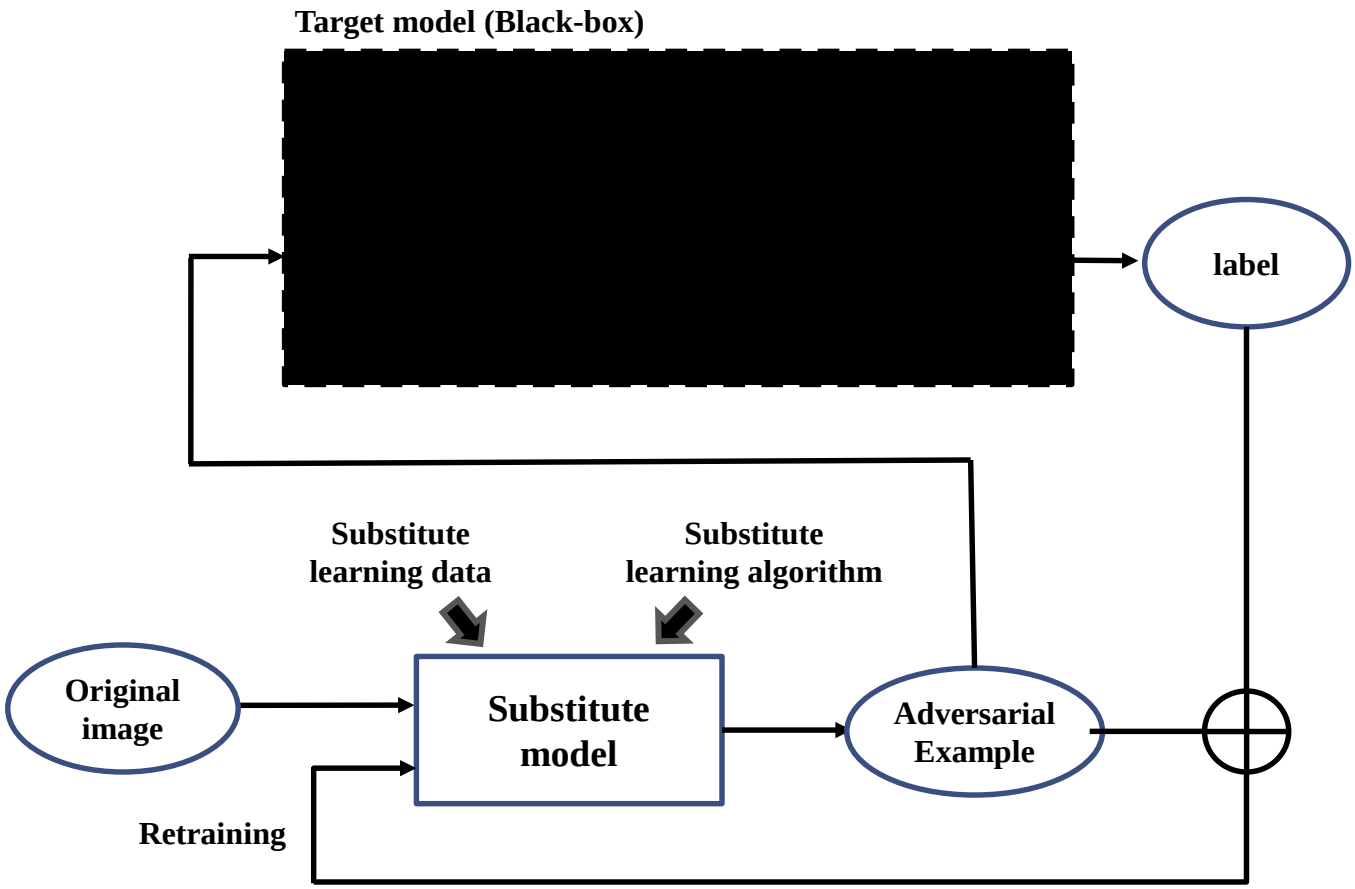
Retraining Substitute Model

▶ 컨셉



Retraining Substitute Model

공격 과정



Issues

- 각자 큰 단점 ➔ 해결
  - 블랙박스 조건 미흡 (confidence, training data)
  - 너무 많은 쿼리 필요
- 공격 성공률 부족
  - 블랙박스 조건 상정
  - 큰 데이터셋 상정 (이미지 크기, class 수)
- Image classification 한정
  - 아직 다른 분야에서는 연구 부족