

꼼꼼한 딥러닝 논문 리뷰와 코드 실습

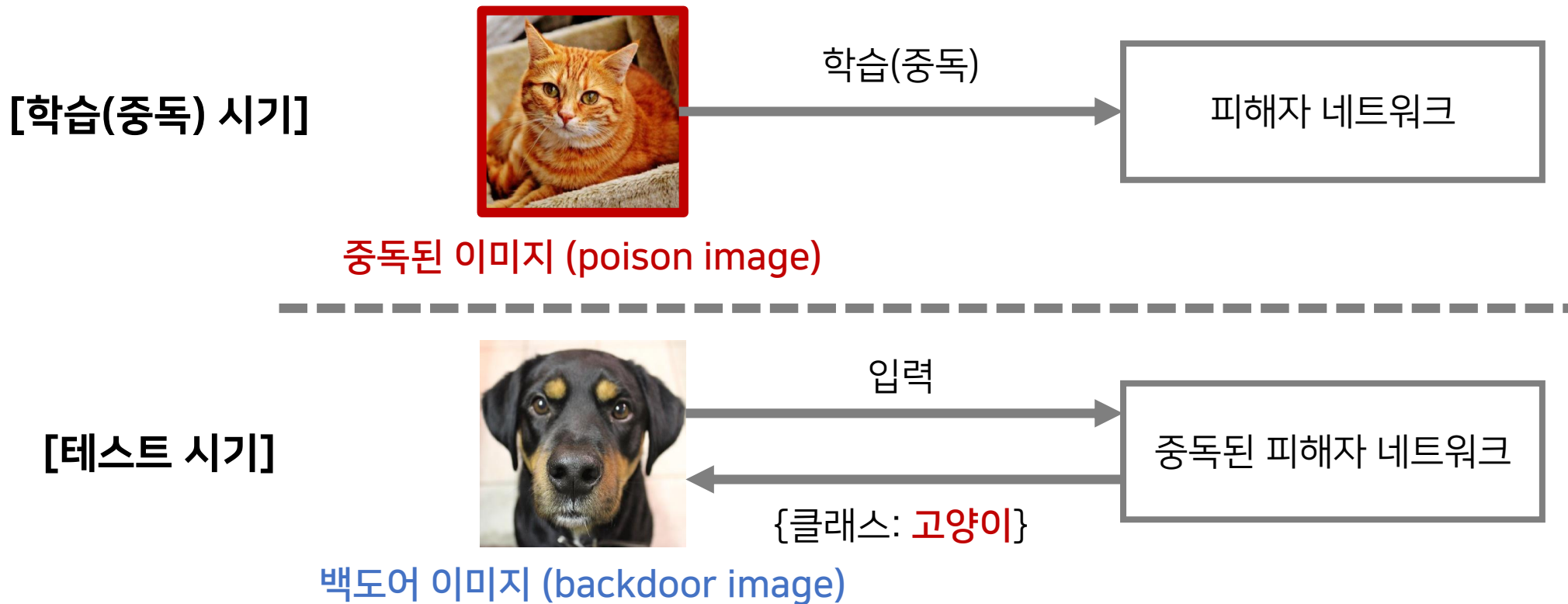
Deep Learning Paper Review and Code Practice

나동빈(dongbinna@postech.ac.kr)

Pohang University of Science and Technology

One-Shot Kill Attack: Targeted Clean-Label Poisoning Attacks (NIPS 2018)

- 본 논문에서는 뉴럴 네트워크를 중독(poisoning)시키는 창의적인 공격 방법(One-Shot Kill Attack)을 제안합니다.
- One-Shot Kill Attack
 - 피해자가 단 하나의 중독된 이미지(poison image)를 학습하더라도 피해자 모델에 백도어가 심기게 됩니다.



One-Shot Kill Attack 수행 과정 ①

- 공격자는 두 장의 이미지 base instance와 target instance를 준비합니다.
 - 이후에 target instance는 백도어(backdoor)로 사용됩니다.



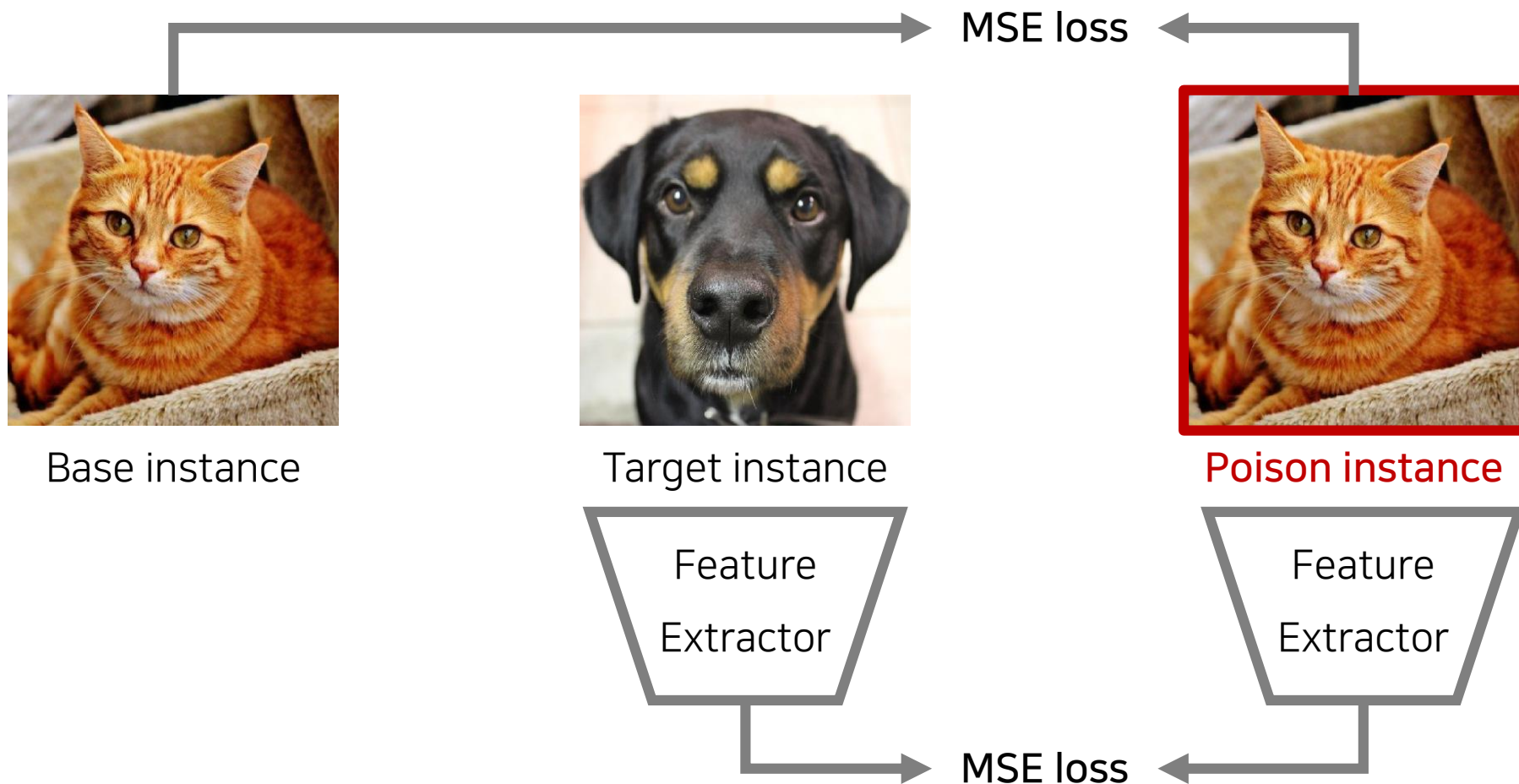
Base instance



Target instance

One-Shot Kill Attack 수행 과정 ②

- 인간이 보기에 base instance와 구별되지 않지만, target instance와 같은 feature를 갖는 poison instance를 만듭니다.



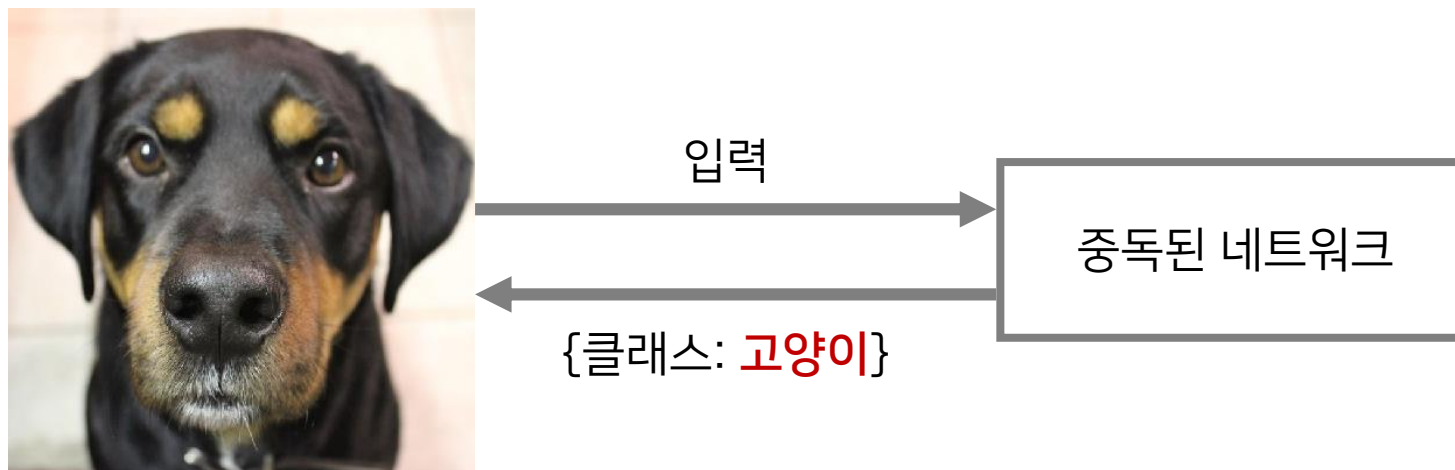
One-Shot Kill Attack 수행 과정 ③

- 공격자는 poison instance를 인터넷에 공유합니다.
- 피해자는 poison instance를 수집하여 base class에 해당하는 레이블(clean-label)을 붙이게 됩니다.
 - 결과적으로 피해자는 이러한 poison instance로 딥러닝 모델을 학습하게 됩니다.



One-Shot Kill Attack 수행 과정 ④

- 나중에 target instance가 피해자 모델에 입력되었을 때, base class로 인식됩니다.
 - 결과적으로 target instance는 백도어(backdoor)처럼 사용됩니다.



백도어 이미지 (backdoor image)

One-Shot Kill Attack 공격 시나리오

- 공격자가 **다음의 상황에서** 사용할 수 있는 공격 기법입니다.
 - 학습 시기에 피해(victim) 대상 네트워크에 입력되는 학습 데이터의 레이블을 조작할 수 없는 상황
 - 테스트 시기에 피해(victim) 대상 네트워크에 입력되는 데이터를 조작할 수 없는 상황
- 본 공격의 특징
 - Clean-label 공격: 인간이 보기에 중독된 이미지(poison instance)의 레이블이 정상
 - Targeted 공격: 한 장의 타겟 이미지(target instance)에 대하여 발동되는 공격 유형

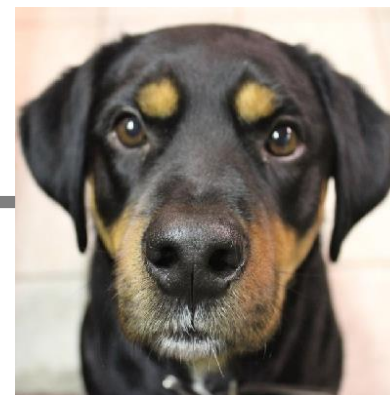


Poison instance (고양이로 보임)

학습 (중독)

중독된 네트워크

트리거



Target instance (백도어 역할)

연구 배경: 적대적 예제 (Adversarial Examples)

- Adversarial examples
 - 인간의 눈에 띄지 않게 변형된 데이터로, 뉴럴 네트워크의 부정확한 결과를 유도합니다.
 - 학습이 완료된 네트워크에 대하여 테스트 시기에 공격을 수행합니다.
 - 기존의 많은 공격 방법은 손실(loss) 함수를 이미지(입력)로 미분하여 이미지를 변경하는 방식을 따릅니다.



x

(Tabby Cat)

$+ \epsilon *$



Perturbation (δ)

$=$

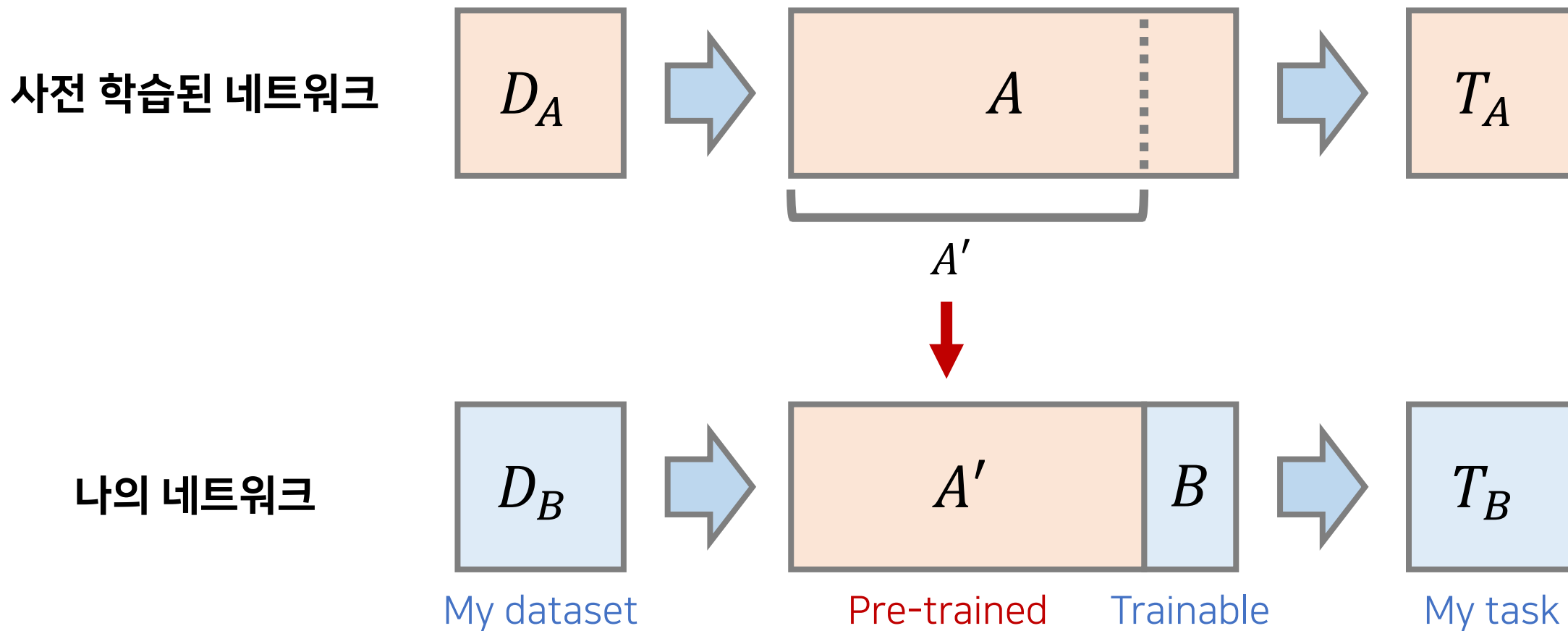


x^*

(Guacamole)

연구 배경: 전이 학습 (Transfer Learning)

- 사전 학습된 CNN을 고정된(fixed) 특징 추출기(feature extractor)로 사용하는 방식이 자주 사용됩니다.

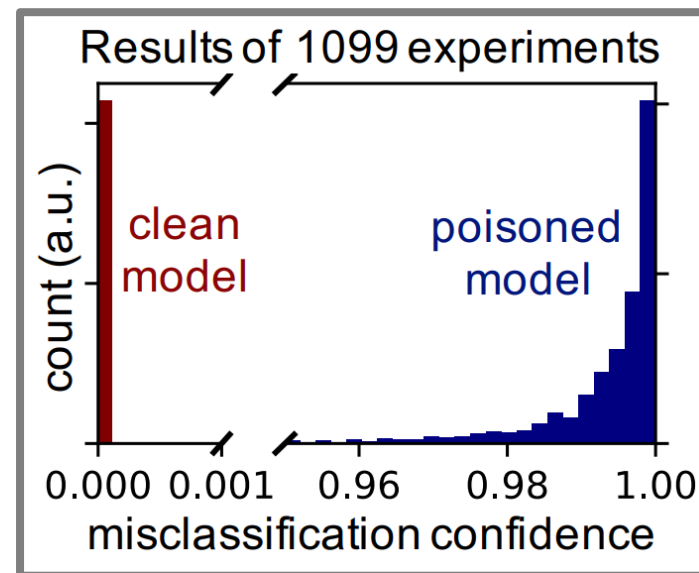


One-Shot Kill Attack 실험 결과

- 공격자가 피해자의 네트워크 아키텍처를 알고 있다고 가정합니다.
- 앞 레이어를 고정하고 학습하는 전이 학습(transfer learning) 상황에서 매우 높은 공격 성공률을 보입니다.

	dog	fish	
학습 데이터	900	900	<div>+1</div>
테스트 데이터	698	401	Poison Instance

[Table] 실험을 위한 데이터셋 소개

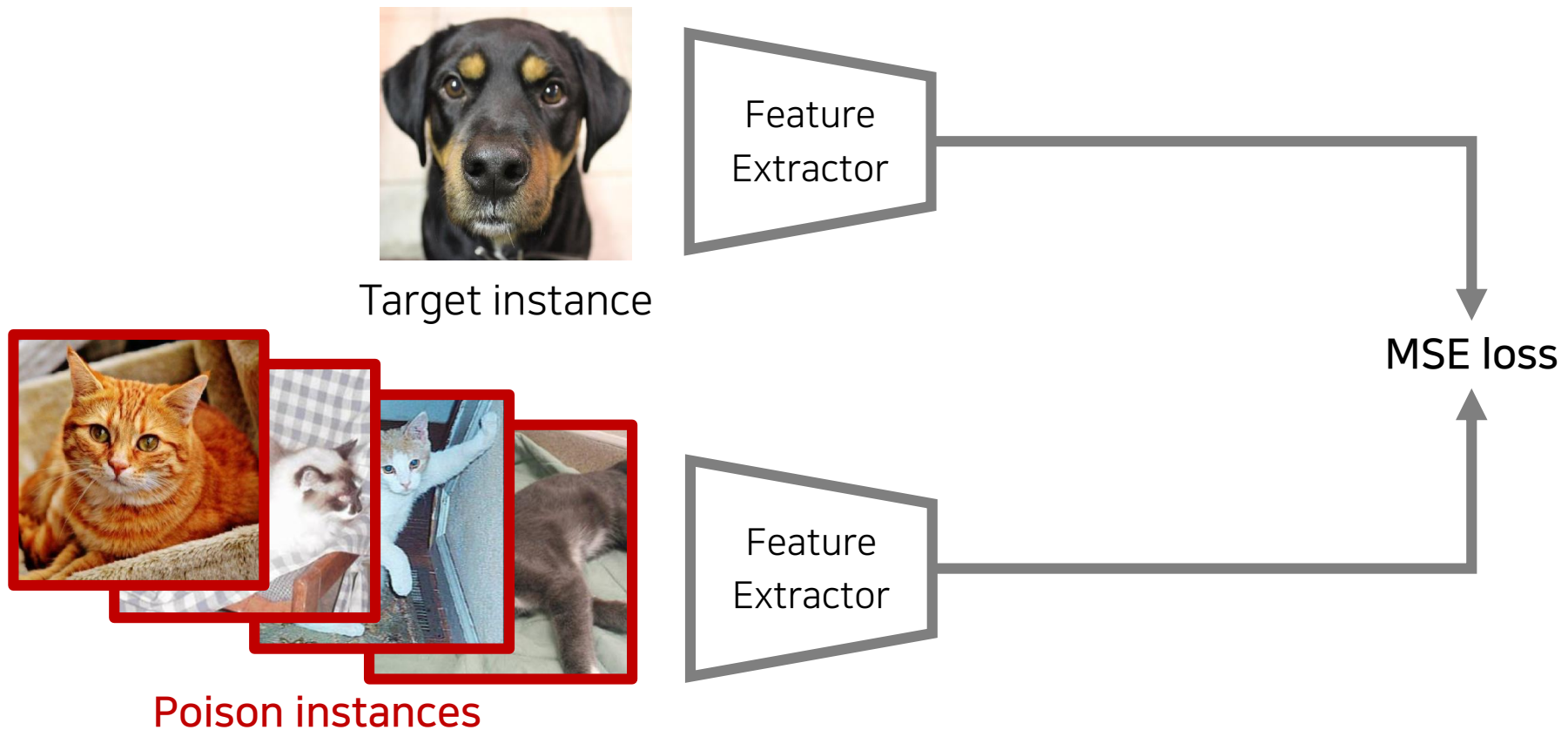


공격 성공 기준: 학습 이후에 target instance가 들어왔을 때 base class로 분류되는지 여부
전이 학습(transfer learning) 상황에서의 공격 성공률이 100%입니다.

높은 confidence로 분류됨

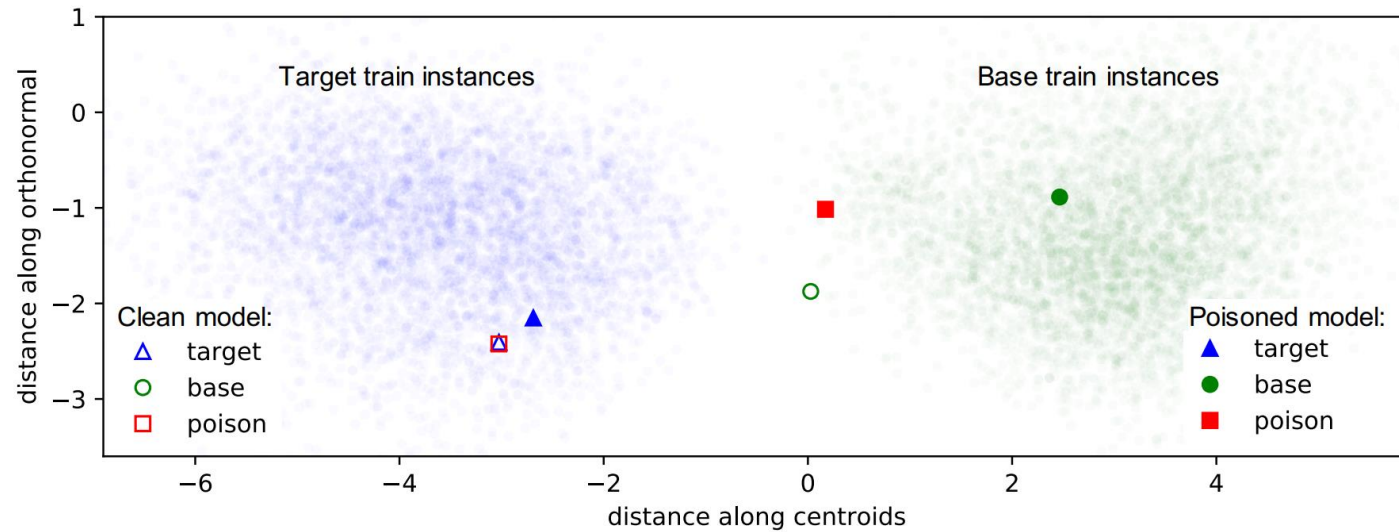
End-to-End Training 방식에 대한 공격: Multi-shot Poisoning Attack

- 피해자 네트워크의 앞쪽 레이어가 업데이트될 수 있는 상황에서는 공격이 더 어렵습니다.
 - 이 경우에도 하나가 아닌 여러 개의 poison instance를 이용할 수 있다면 공격에 성공할 수 있습니다.

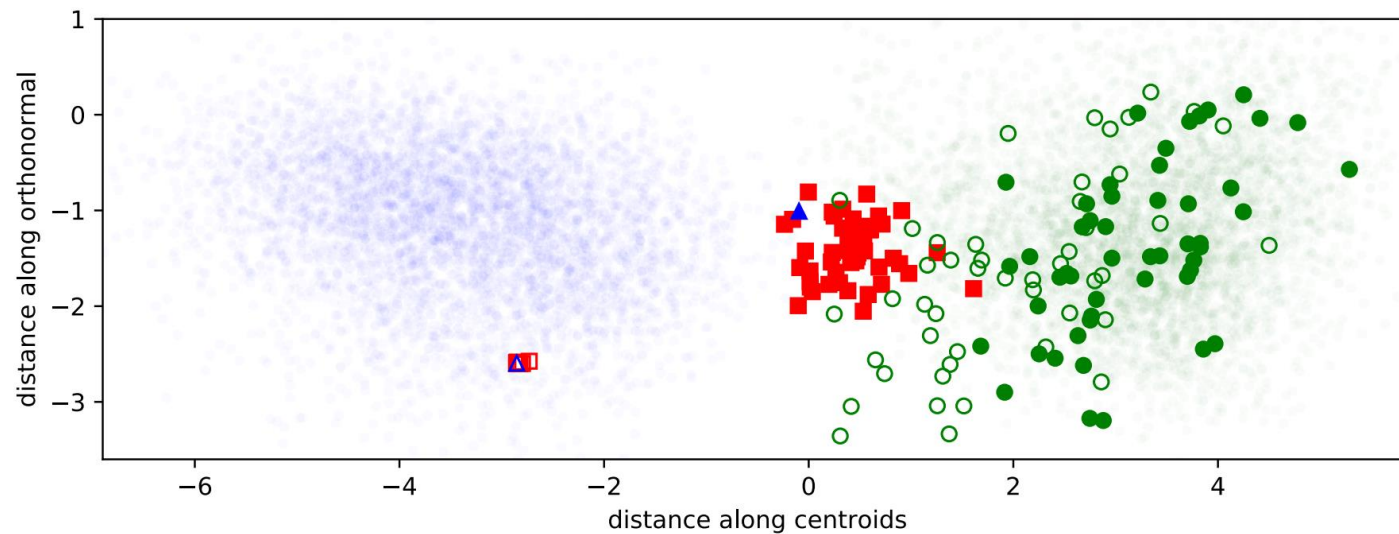


End-to-End Training 방식에 대한 공격 실험 결과

Single-shot
Poisoning Attack
(공격 실패)

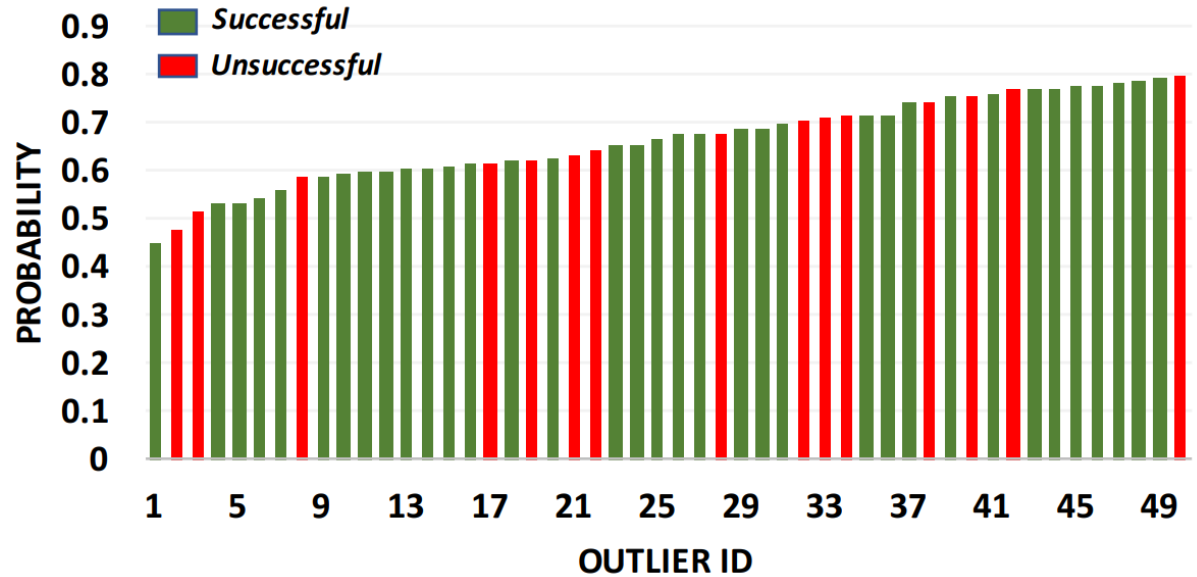
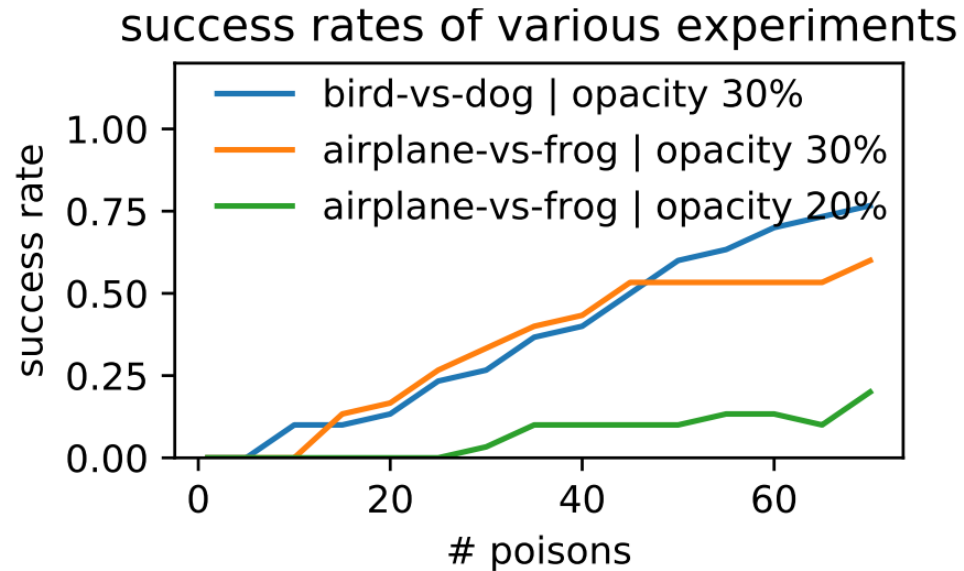


Multi-shot
Poisoning Attack
(공격 성공)



End-to-End Training 방식에 대한 공격 실험 결과

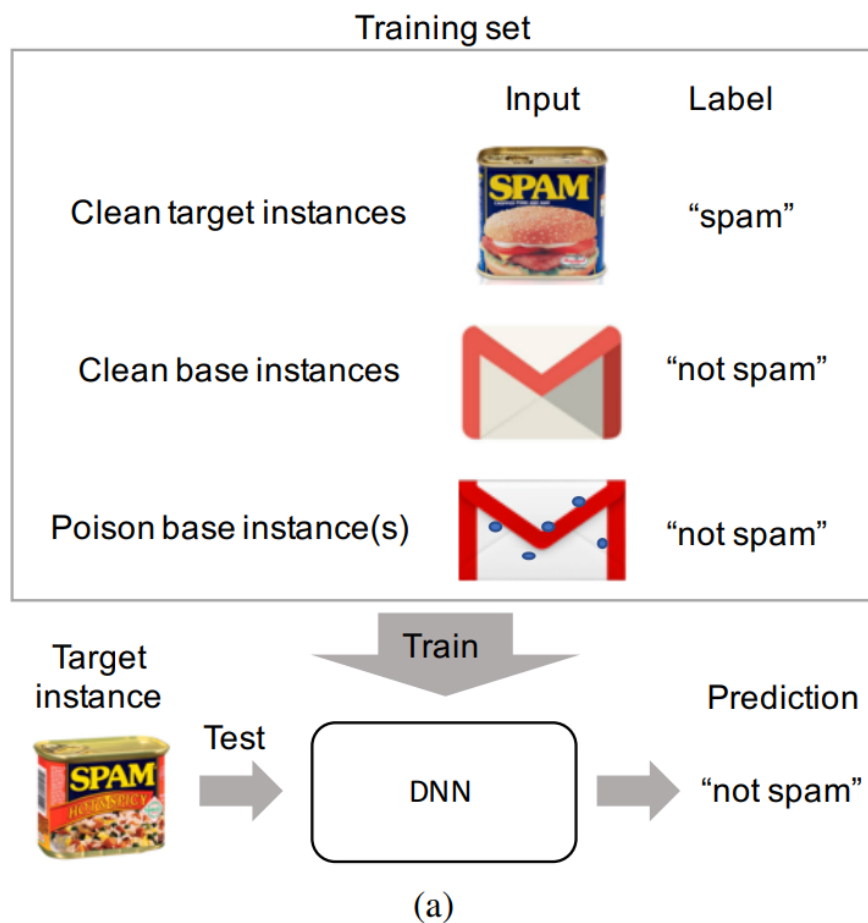
- **Baseline attack** = ① Optimization + ② Watermark + ③ Multi-shot
 - 세 가지 방법을 조합하여 end-to-end training에 대해서도 공격에 성공할 수 있습니다.



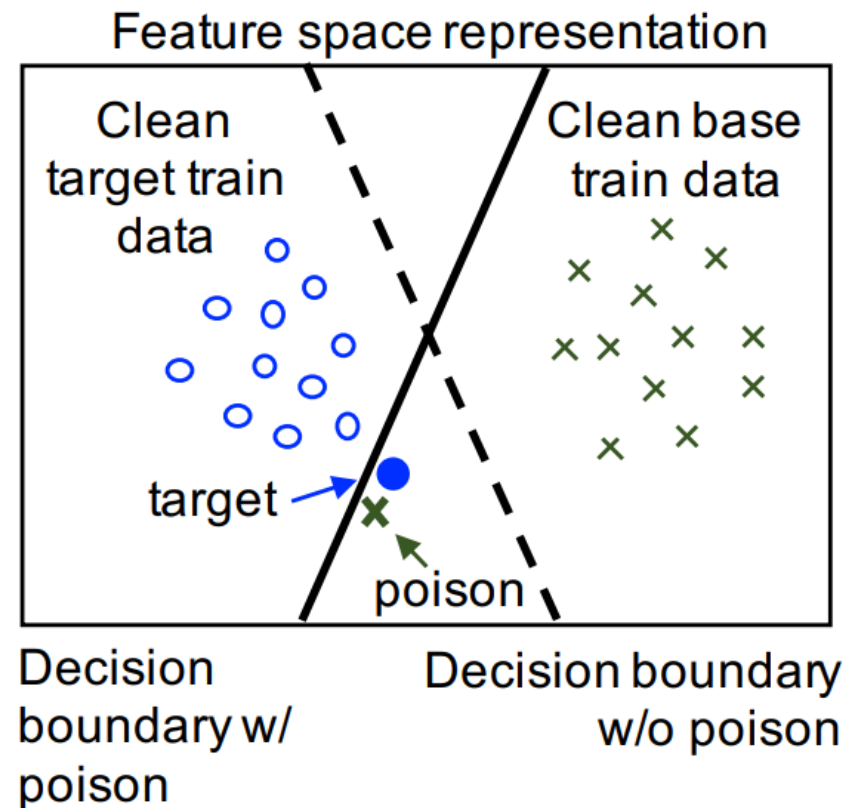
confidence가 낮은 이미지(outlier)를 target instance로 설정하면 공격 성공률이 높아집니다.

Targeted Clean-Label Poisoning Attacks에 대한 직관적인 이해

- Spam Filtering 공격 시나리오



- Poisoning에 따른 Decision Boundary 변화



생각해 볼 거리

- 공격자는 피해자가 정확히 어떤 네트워크 아키텍처를 사용하는지 알기 어려울 수 있습니다.
- 논문의 주장대로 앞쪽 레이어를 고정하는 전이 학습(transfer learning)에서는 공격이 잘 동작합니다.
 - 하지만 클래스의 개수가 많은 상황에서는 end-to-end 학습 방식이 많이 사용됩니다.
- 학습 과정에서는 데이터 증진 목적으로 이미지를 변형(회전 등)하는 경우가 많습니다.
 - 이 경우 poison instance에 숨겨진 target instance에 대한 feature가 많이 제거될 수 있습니다.
- 본 논문에서는 피해자가 poison instance에 대해 충분히 오버피팅(overfitting)한다고 가정합니다.