캡 스 톤 디 자 인 I 최 종 발 표

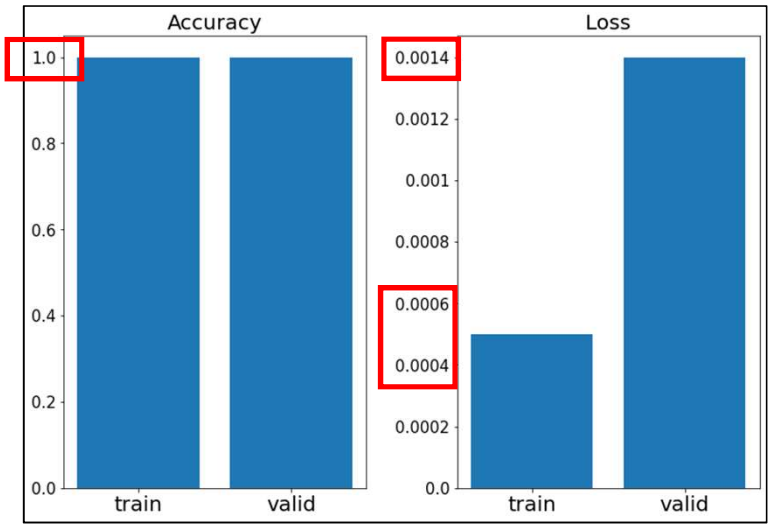
안티포렌식에 강인한 딥페이크 탐지 기법

민지민, 김지수, 김민지

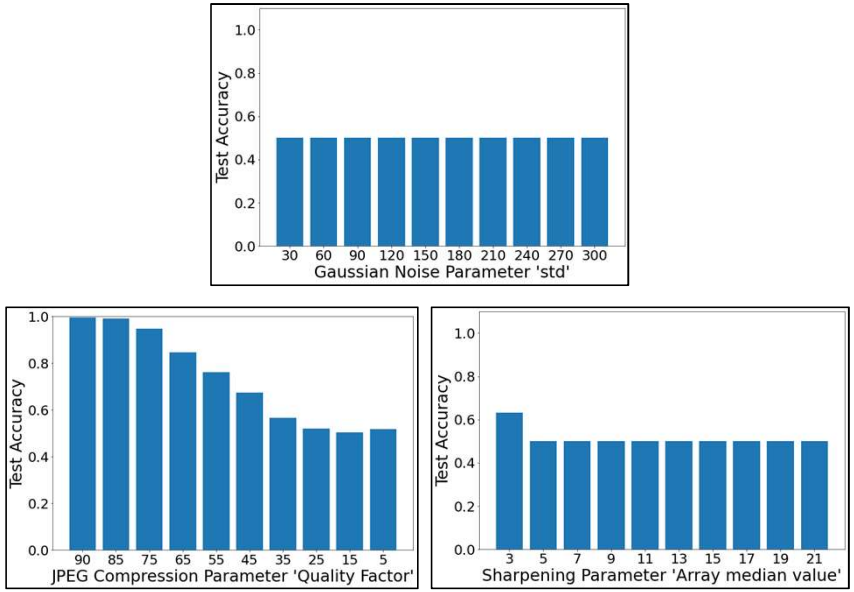
개발 배경 및 목적

기존 딥페이크 탐지 모델의 문제점

- **간단한 이미지 편집**(ex. Sharpening, additive noise) 적용한 경우 **탐지 모델 -> 무력화**



원본 데이터셋 학습 모델 성능



편집된 이미지에 대한 정확도

탐지 기술 무력화 -> 안티 포렌식

적대적 데이터셋을 생성하는 안티 포렌식 공격

- 화이트 박스 공격(White-Box Attacks)
  - 공격자가 **탐지 모델에 대한 정보를 모두 안다**는 전제
  - 비현실적 조건
  - 공격 성공률 100% 가까움

- 블랙 박스 공격(Black-Box Attacks)
  - 공격자가 **원본 이미지에 특정 노이즈 추가 -> 오분류 유도**
  - 실제 화이트 박스 공격보다 많이 시도됨

이미지 출처: https://sigport.org/documents/countering-anti-forensics-image-re-sampling#files

두 가지의 블랙 박스 공격

- **Adversarial Attack**
  - 데이터셋에 네트워크를 교란시키는 노이즈를 추가
  - **초보자가 쉽게 공격할 수 있는 방법 X**

- **이미지 편집 기법**
  - 데이터셋에 이미지 편집을 가함
  - **초보자가 쉽게 공격할 수 있는 방법 O**

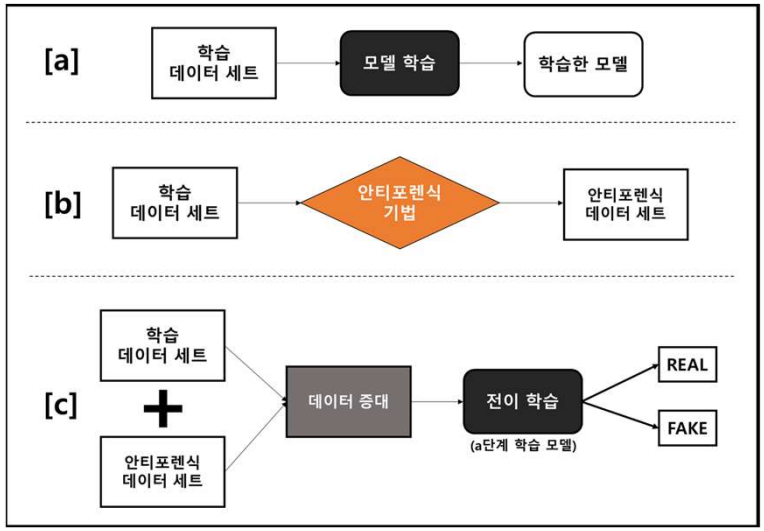
⇒ **두 방법 모두 딥페이크 탐지기 무력화 가능**

전문가가 아니더라도 공격할 수 있는
이미지 편집 기법 기반의 블랙 박스 공격에 대응할 기술 필요

이미지 편집 기법을 이용한 블랙 박스 공격에 대응

=> 적대적 학습 수행

<탐지 모델 학습 절차>



학습 데이터셋에 안티 포렌식 데이터셋 추가
-> 공격 패턴 학습하여 강인한 모델 생성

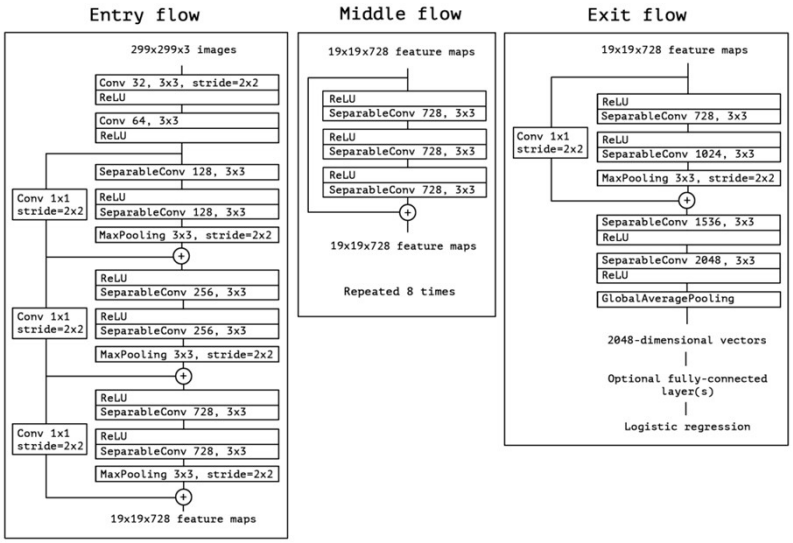
직접 구축한 안티 포렌식 데이터셋을 학습 데이터셋에 추가 -> 적대적 학습
-> 강인한 모델 개발

이미지 출처: https://towardsdatascience.com/adversarial-machine-learning-mitigation-adversarial-learning-9ae04133c137

학습 네트워크 구조

Xception 네트워크

-> 경험적으로 좋은 성능을 보이는 네트워크를 기반으로 진행하고자 함



네트워크 구조

[딥페이크 변조영상에 대한 탐지율]

Accuracies	DF	F2F	FS	NT	Real	Total
Xcept. Full Image	74.55	75.91	70.87	73.33	51.00	62.40
Steg. Features	73.64	73.72	68.93	63.33	34.00	51.80
Cozzolino *et al.*	85.45	67.88	73.79	78.00	34.40	55.20
Rahmouni *et al.*	85.45	64.23	56.31	60.07	50.00	58.10
Bayar and Stamm	84.55	73.72	82.52	70.67	46.20	61.60
MesoNet	87.27	56.20	61.17	40.67	**72.60**	66.00
XceptionNet	**96.36**	**86.86**	**90.29**	**80.67**	52.40	**70.10**

출처: Rossler, Andreas, et al. "**Faceforensics++**: Learning to detect manipulated facial images." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

네트워크 구조 이미지 출처: https://openaccess.thecvf.com/content_cvpr_2017/html/Chollet_Xception_Deep_Learning_CVPR_2017_paper.html

실험 데이터셋 구축

안티 포렌식 데이터셋 생성

안티 포렌식 공격으로 적용한 이미지 편집 기법
-> Sharpening, Gaussian Noise, JPEG Compression

원본 fake 이미지

Sharpening이 적용된 fake 이미지

Why use?

딥페이크 변조 이미지는 합성된 부분에서 **blurry 특징**
-> 위 이미지 편집을 통해 **fake 이미지의 blurry 특징을 감쇄**

실험 데이터셋 구축

안티 포렌식 데이터셋 생성 시 적용 사항

- **실제 발생할 수 있는 공격 상황은 예측 불가**
  -> 공격(이미지 편집 기법) 별 강도 10단계 설정

- **적용할 3가지 공격 강도의 수준은 PSNR 사용해 조절**
  -> PSNR 30 이하의 이미지는 시각적 품질이 많이 떨어져 사용 X

- **안티 포렌식 공격 목표: real -> fake로 판별, fake -> real로 판별**
  -> 학습 데이터셋(real & fake) 전부 이미지 편집 적용

실험 데이터셋 구축

안티 포렌식 데이터셋 생성

Sharpening

Attack Level	Sharpening array 중앙값
1	3
2	5
3	7
4	9
5	11
6	13
7	15
8	17
9	19
10	21

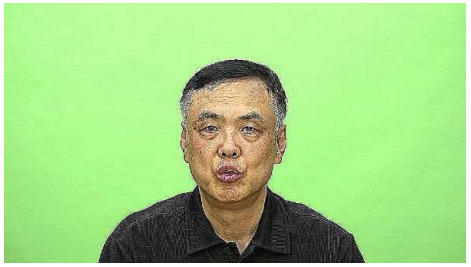
적용 강도

중앙값 = **3**

중앙값 = **9**

중앙값 = **15**

중앙값 = **21**



예시 이미지
(real 이미지에 적용)

실험 데이터셋 구축

안티 포렌식 데이터셋 생성

Gaussian Noise

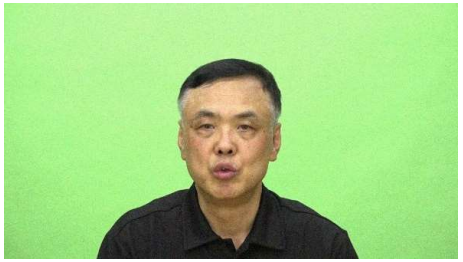
Attack Level	Standard deviation
1	30
2	60
3	90
4	120
5	150
6	180
7	210
8	240
9	270
10	300

적용 강도

Std = **30**

Std = **120**

Std = **210**



Std = **300**

예시 이미지
(real 이미지에 적용)

실험 데이터셋 구축

안티 포렌식 데이터셋 생성

JPEG Compression

Attack Level	Quality Factor
1	90
2	85
3	75
4	65
5	55
6	45
7	35
8	25
9	15
10	5

적용 강도

QF = **90**

QF = **65**

QF = **35**

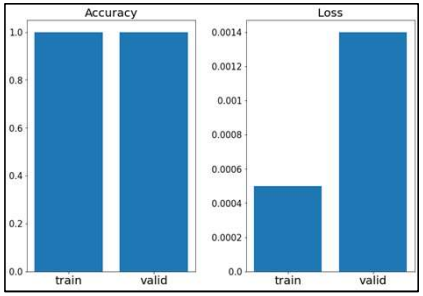
QF = **5**

예시 이미지
(real 이미지에 적용)

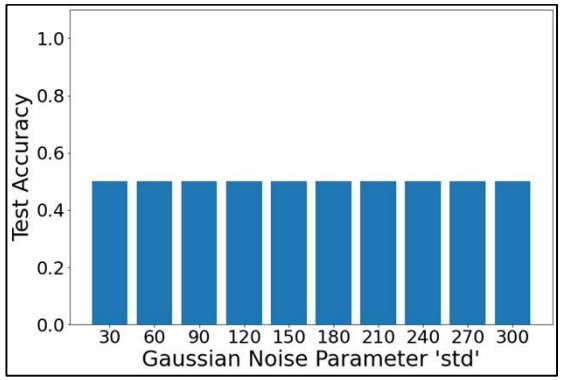
안티 포렌식 공격에 대한 정확도 하락 확인

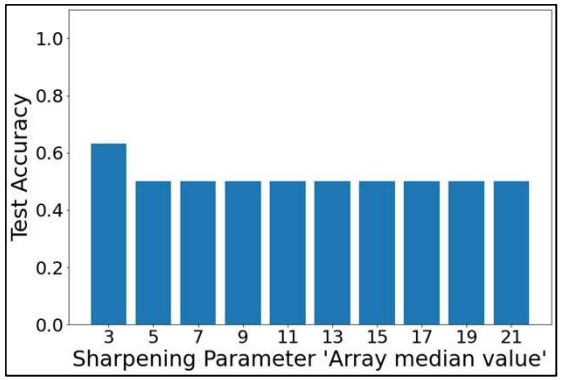
안티 포렌식 데이터셋에 대한 성능 하락 확인

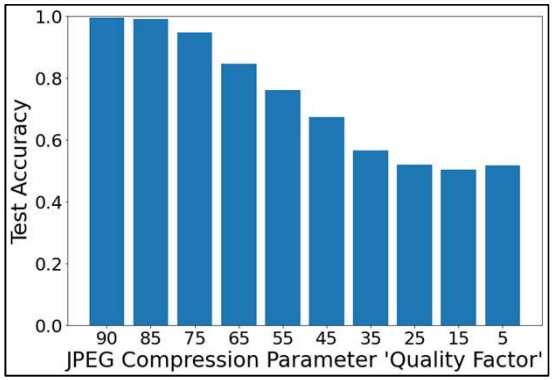
-> 원본 데이터셋만 학습한 모델 사용



원본 데이터셋 학습 모델 성능



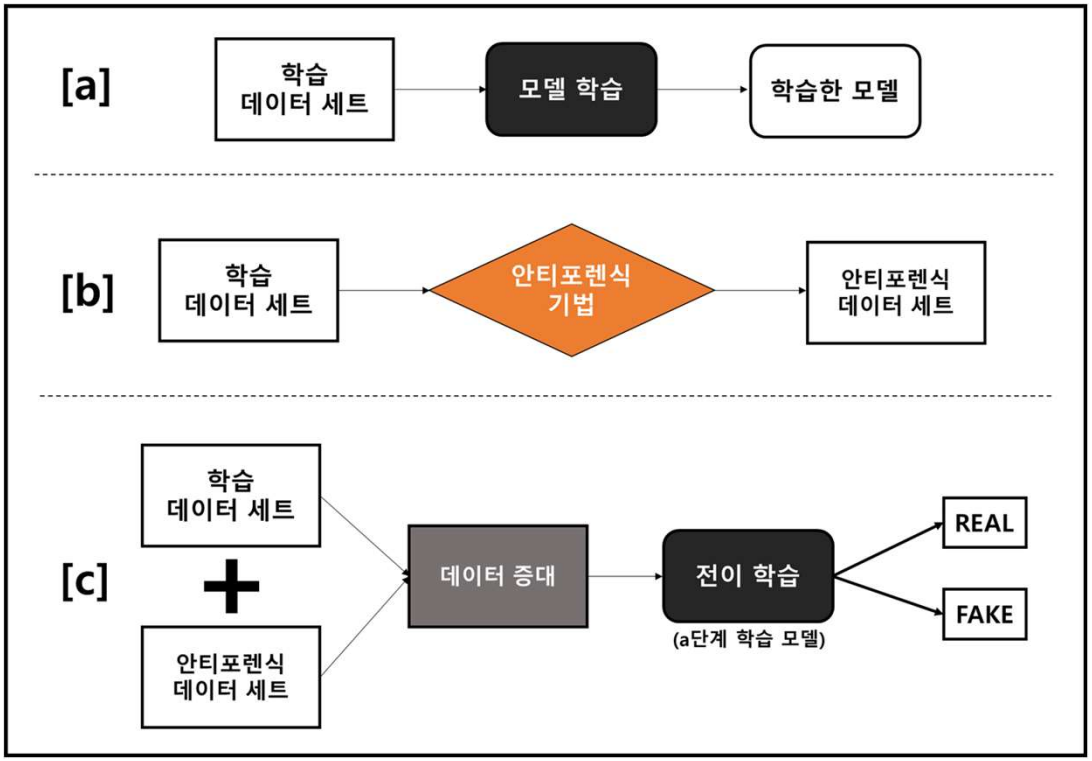




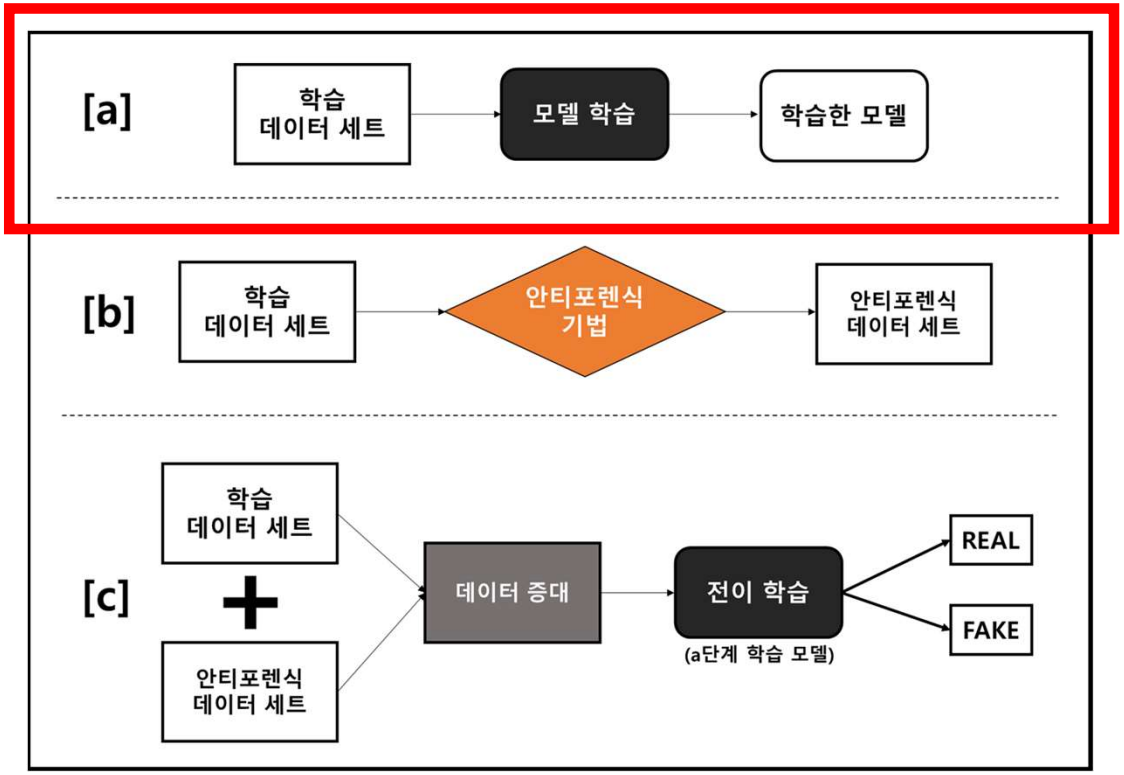
공격 강도 별 탐지 모델 **성능 매우 하락**

=> 탐지 모델이 안티 포렌식에 취약

제안하는 적대적 학습 기법



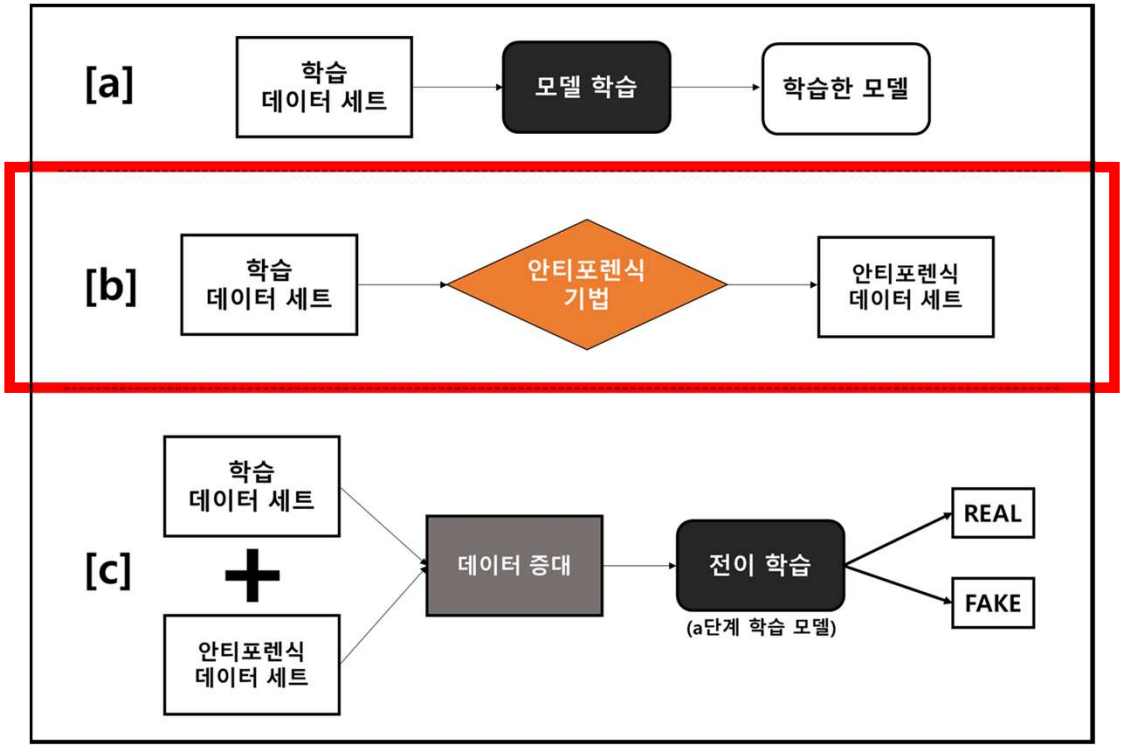
제안하는 적대적 학습 기법 – 1단계



<원본 데이터셋 학습>

원본 데이터셋만 사용해
Xception 네트워크 학습
-> 모델 저장

제안하는 적대적 학습 기법 – 2단계

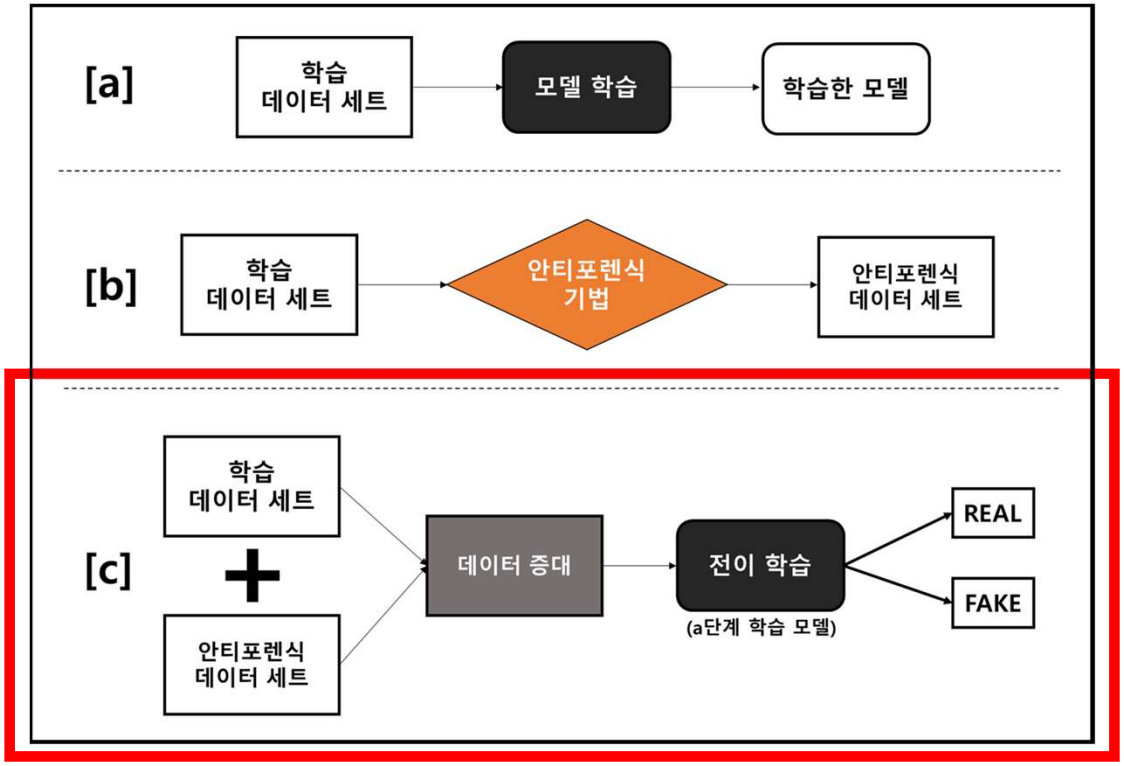


<안티 포렌식 데이터셋 생성>

3가지 이미지 편집 기법
각각 10단계의 강도로 적용

=> 30종의 안티 포렌식 데이터셋 생성

제안하는 적대적 학습 기법 – 3단계



<전이학습을 활용한 적대적 학습 수행>

[a]에서 도출된 학습 모델의
전이 학습에 사용할 학습 데이터셋
-> 원본 데이터셋 + 안티 포렌식 데이터셋
⇒ 모델 생성

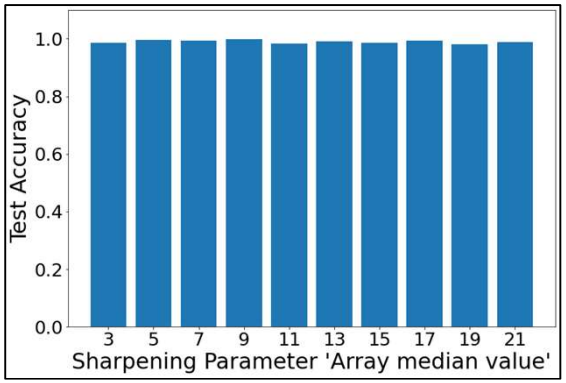
ex) 원본 데이터셋 + sharpening Lv1 데이터셋
=> **Sharpening Lv1 공격에 강인한 모델 생성**

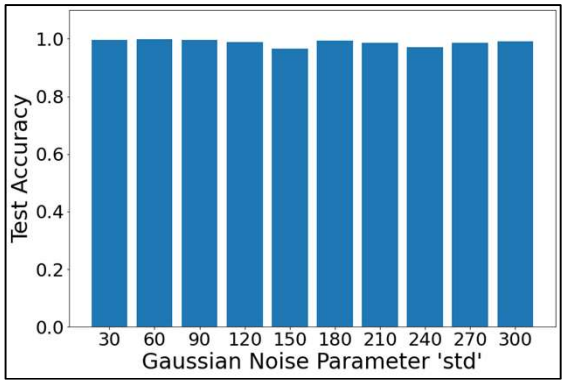
하이퍼파라미터 설정

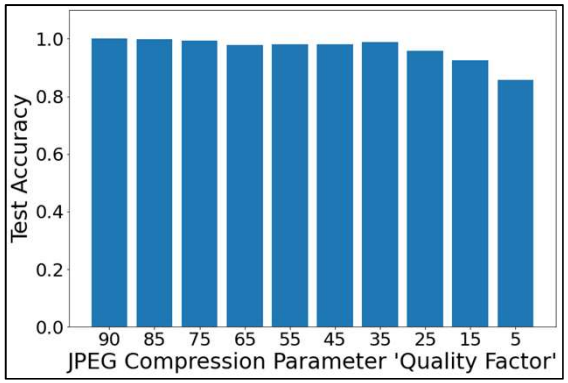
Epoch	**3**
Batch size	16
Optimizer	Adam
Learning rate	0.001 (초기값)
LR Scheduler	StepLR
Loss	CrossEntropyLoss

적대적 학습 적용 모델의 성능

테스트셋 – 학습에 적용한 공격&강도의 안티 포렌식 데이터셋으로만 구성







거의 모든 경우의 탐지 모델이
공격이 가해진 데이터셋에 대해 높은 정확도 보임

결론

기존: 탐지 정확도가 높은 모델이라도 **간단한 이미지 변형으로 탐지 우회** 가능

적대적 학습 기법 적용 -> 안티 포렌식 공격 패턴 학습

안티 포렌식 데이터셋에 대해 높은 탐지 강인성 획득 가능

팀원의 역할

✓ 각자 주어진 환경에 맞추어 업무 분담
✓ 학술대회 논문은 함께 작성

김지수	김민지	민지민
실험 데이터셋 생성	학습 모델 생성	학습 모델로 성능 확인 (inference)

우수성 입증 자료

✓2022 정보처리학회 '학부생 논문경진대회' 동상 수상

• ASK 2022(춘계) 학부생논문경진대회 수상자 (31명)

포상명	수상자	소 속	공동저자
대 상 (1)	김동원	성균관대학교	김동원, 신유정
금 상 (2)	마상균	영남대학교	마상균, 박재현, 서영석
	강민주	숙명여자대학교	강민주, 전우진, 윤용익
은 상 (5)	오은비	서울여자대학교	강채희, 오은비, 이승언, 이현경, 김성욱
	최욱철	성균관대학교	최욱철, 구자환, 김응모
	우성미	중앙대학교	허정회, 우성미, 이대원
	한재상	숭실대학교	한재상, 강윤서, 권채연, 허원제, 김영종,
	송성호	경기대학교	송성호, 김인철
동 상 (8)	장준보	한국외국어대학교	장준보
	이계혁	홍익대학교	이계혁, 황민채, 구영인, 현동엽, 유동영
	양수빈	상명대학교	양수빈, 김민태, 권수빈, 우나현, 김학재, 정태경, 이성주
	민지민	한발대학교	민지민, 김지수, 김민지, 장한얼
	전기범	숭실대학교	전기범, 이유진, 안민하, 김용규, 김영종
	정승균	대구가톨릭대학교	정승균, 김규동, 김병창
	박정현	호서대학교	박정현, 송민석, 백찬영, 홍우성, 김언정, 문남미
	문지원	대구가톨릭대학교	문지원, 김강산, 김영은, 강다은, 이제민, 황정민, 이제원, 김동주
장려상 (15)	한성민	국립한경대학교	한성민, 박명숙, 김상훈
	이연아	홍익대학교	이연아, 정유진, 하지혜, 이수연, 유동영
	Edward Dwijayanto Cahyadi	세명대학교	Edward Dwijayanto Cahyadi, 송미화
	정해린	동덕여자대학교	정해린, 김도영, 정현정, 김성경, 김현희
	박선호	숭실대학교	박선호, 박근형, 김태현, 김유림, 김영종
	박병수	숭실대학교	박병수, 박진혁, 임성현, 유준열, 김영종
	장환곤	숭실대학교	장환곤, 박성철, 나상우, 김 민, 이영제, 김영종
	김나연	동덕여자대학교	김나연, 김도영, 김미려, 정지영, 김현희
	신준석	서원대학교	신준석, 이덕규
	채수지	광운대학교	권나은, 이규민, 이지윤, 채수지, 박규동, 이상민
	박규환	서원대학교	박규환, 이덕규
	이창열	한양대학교 ERICA	이창열, 이진현, 차정현, 서승현
	박영서	백석대학교	박영서, 강 혁, 이근호
	이종환	세명대학교	이종환, 곽희웅, 박기수, 송미화
	차건환	숭실대학교	차건환, 김희수, 박정언, 박성준, 김영종

제 2022-6-12호

동 상

민지민, 김지수, 김민지, 장한얼(한발대학교)

제목 : 안티 포렌식에 강인한 딥페이크 탐지 기법

귀하께서는 본 학회가 주최하는 ASK 2022의 논문경진대회에 논문을 제출하여 심사위원단의 엄정한 심의를 거쳐 동상으로 선정되었기에 이 상을 수여합니다.

2022년 5월 20일

사단법인 한국정보처리학회
회 장 강 진 모

질의 응답