작품(과제)명 : 스테그어날리시스 성능 개선 모델

팀명 : 봇듀오

참여학생 : 김동수 오서연

지도교수 : 장한얼

작품 개요

이미지에 악성 코드와 같은 데이터를 삽입하는 스테가노그래피 기술에 대응하기 위해 숨겨진 데이터를 탐지하는 스테그어날리시스 기술이 사용되고 있다. 하지만, 숨겨진 데이터는 굉장히 미세해서 육안으로는 확인하기 어렵고, 커버 데이터의 촬영 환경과 같은 도메인 정보에 굉장히 민감하게 반응하여 성능에 큰 영향을 끼친다.

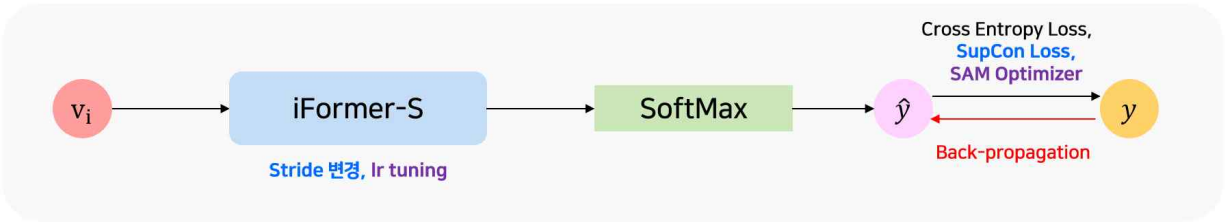
이를 해결하기 위해 최적의 Backbone 모델과 학습 방법을 탐색하고, 가장 우수한 성능을 보이는 방식을 선정하였다. 또한, 학습에 사용된 기기의 데이터뿐만 아니라 다른 기기에서 생성된 데이터로도 성능을 평가하여, 실제 환경에서의 스테그어날리시스 성능을 개선한다.

작품 추진과정 사진

- 실험에 사용한 기기는 다음과 같다.
  - ■ 공간 영역(PNG): 총 14개 기기의 PNG 포맷 데이터 사용
  - ■ 압축 영역(JPEG): 총 19개 기기의 JPEG 포맷 데이터 사용

[그림1] 학습에 사용한 스마트 기기 데이터셋

- 가장 높은 정확도를 보여준 iFormer-Small 모델을 Backbone 모델로 사용하였다.
- 단일 기기로 학습하였을 때보다 기기 통합 데이터로 학습 시 전체적으로 좋은 성능을 보여주었다.
- 적용 메소드는 **Stride 변경**, Sharpness-Aware Minimization(**SAM**), Supervised Contrastive Learning (**SupCon**)으로, 기존 모델의 Stride를 (2,2)에서 (1,1)로 변경해 이미지의 Downsampling을 줄여 정보 손실을 최소화한다. SAM Optimizer는 손실 함수의 평평한 영역에서 최솟값을 찾아 모델의 일반화 성능을 개선하고, SupCon Loss는 같은 클래스의 데이터는 가까워지게, 다른 클래스의 데이터는 멀어지도록 하여 모델의 일반화 성능을 개선한다.



[그림2] 모델 아키텍처

기대효과 내용

- 선정된 Backbone 모델 및 학습 기법을 사용하여 모델의 일반화 성능을 개선하여 다양한 촬영 환경 변화에도 안정적인 성능을 유지할 수 있다. 이로 인해 숨겨진 악성 코드나 데이터를 더 정확하게 탐지할 수 있어, 실제 환경에서 스테가노그래피 공격에 대한 보안 대응할 수 있다. 이러한 일반화 성능 덕분에 다양한 기기와 플랫폼에서 동일한 성능을 보장해, 다양한 보안 시스템에 폭넓게 활용될 수 있다.

완성품 사진

Device	Method	Train Accuracy	Test Accuracy
Galaxy Flip3	Base	96.91	93.02
	Stride 1	97.16	94.97
Galaxy S20+	Base	95.98	92.32
	Stride 1	97.39	94.34
iPhone12	Base	93.55	88.16
	Stride 1	95.59	93.21
iPhone13 mini	Base	94.14	90.32
	Stride 1	95.4	92.57
Huawei P30	Base	95.98	93.51
	Stride 1	97.41	95.67
LG Wing	Base	95.31	93.12
	Stride 1	97.31	96.77

[표 1] Stride 변형 적용 압축 실험 결과

Device	Method	Train Accuracy	Test Accuracy
Galaxy Flip3	Base	81.69	79.03
	Stride 1	97.13	97.56
Galaxy S20+	Base	49.68	50.32
	Stride 1	85.75	87.91
iPhone12 ProMax	Base	98.94	99.05
	Stride 1	99.71	99.68
Huawei P30	Base	84.08	84.97
	Stride 1	91.2	91.05
LG Wing	Base	50.02	50.2
	Stride 1	50.19	50.13

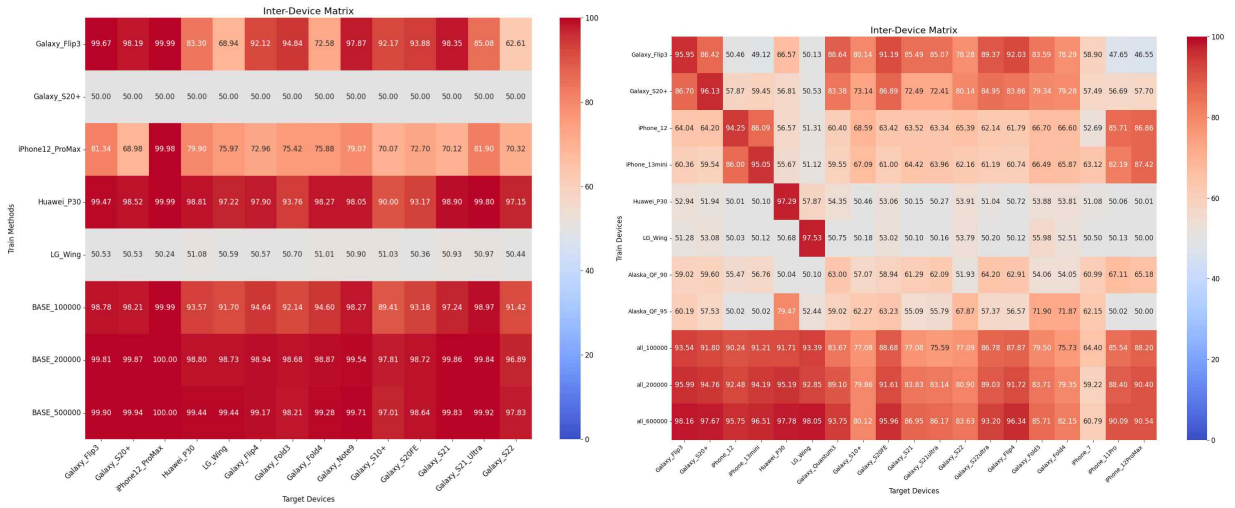
[표 2] Stride 변형 적용 공간 실험 결과

Learning Rate	Device	Train Accuracy	Test Accuracy
1e-5	Galaxy S20+	98.54	98.29
5e-5		99.63	99.75
7e-5		99.75	99.82
1e-5	LG Wing	97.49	96.72
5e-5		49.94	50.25
7e-5		49.94	50.38

[표 3] lr 적용 공간 영역 실험 결과

Learning Rate	Device	Train Accuracy	Test Accuracy
1e-3	Galaxy S20+	50.14	50.05
1e-5		99.14	97.09
5e-4		50.96	50.31
1e-3	LG Wing	91.00	89.79
1e-5		98.1	96.55
5e-4		98.1	98.1

[표 4] lr 적용 압축 영역 실험 결과



[그림 3] 공간·압축 영역 개별 기기 및 통합 데이터셋 실험 결과

Method	Intra Accuracy	Inter Accuracy
Baseline	92.58	88.30
SAM(rho=0.05)	98.16	95.76
7(CE):3(SupCon)	92.85	88.50

[표 1] 공간 영역 통합 데이터셋 실험 결과

Method	Intra Accuracy	Inter Accuracy
Baseline	94.05	83.87
SAM(0.005)	95.10	84.29
7(CE):3(SupCon)	94.51	84.98
SAM(rho=0.001), 7(CE):3(SupCon)	95.05	84.11

[표 2] 압축 영역 통합 데이터셋 실험 결과