RLB-MI 재생산: 강화학습 기반 블랙박스 모델 인버전 공격

CSEG516 강화학습 팀 프로젝트

재생산 연구 강화학습 프라이버시 공격 GAN + ArcFace

팀원: 김대원, 최서빈

서강대학교, 2025년 가을학기

초록 (Abstract): *이 프로젝트는 CVPR 2023의 RLB-MI(강화학습 기반 블랙박스 모델 인버전) 공격 프레임워크를 재생산합니다. 우리는 CelebA 및 FaceScrub 데이터셋으로 훈련된 ArcFace 기반 얼굴 분류기를 대상으로, 사전 훈련된 GAN의 잠재 공간(latent space)을 탐색하기 위해 Soft Actor-Critic (SAC)을 구현했습니다. 추가적으로, 우리는 추론 시* ***로짓 온도 스케일링(inference-time logit temperature scaling)****(ArcFace 스케일 팩터 s 조정)이 공격 성능에 미치는 영향을 조사하여, 보정된(calibrated) 신뢰도 점수가 RL 에이전트에게 더 유익한 보상 신호를 제공한다는 것을 입증했습니다.*

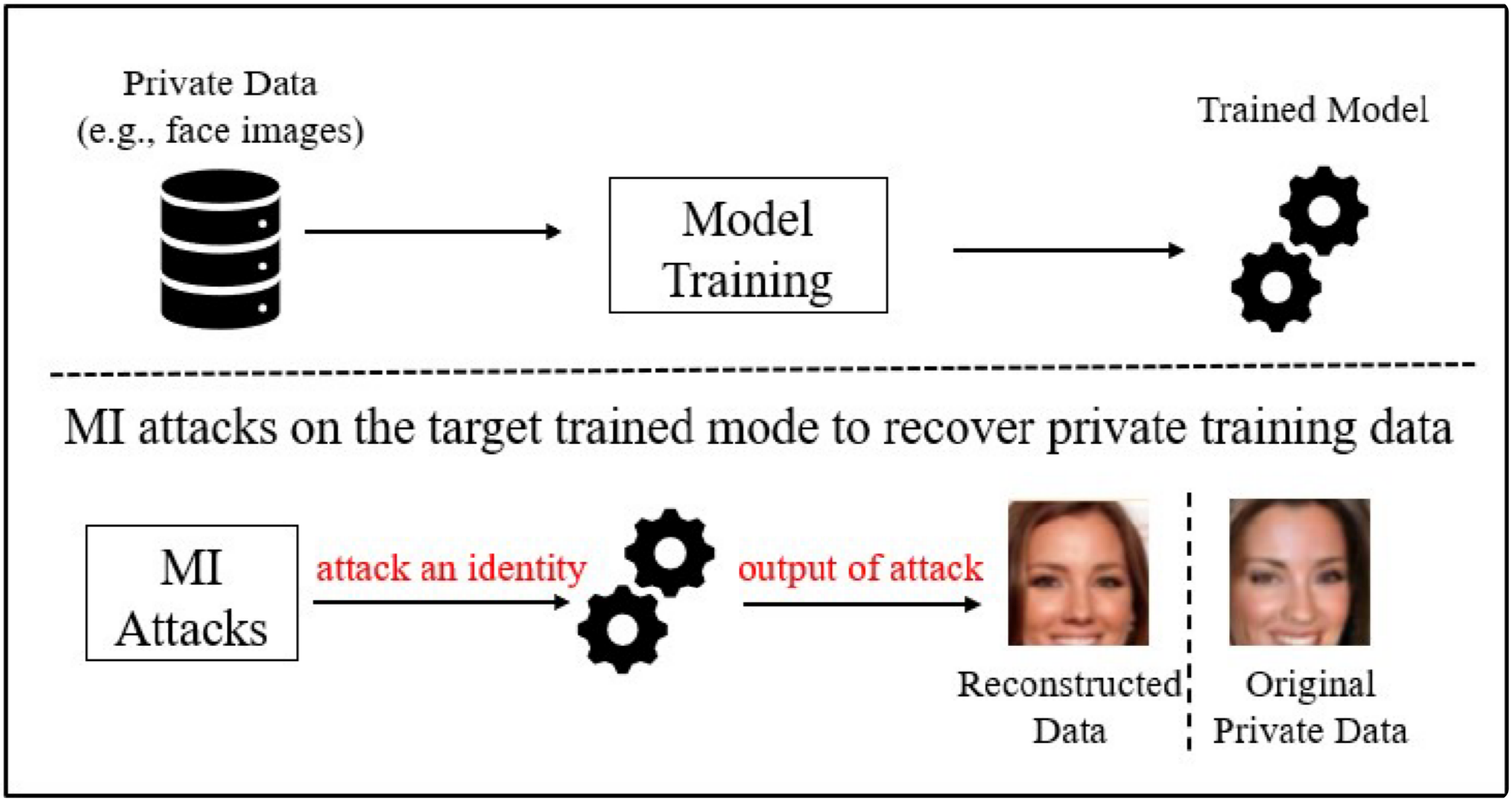

그림: RLB-MI 공격 프레임워크 개요.

1. 서론 (Introduction)

모델 인버전 공격(MIA)은 모델의 출력과 훈련 데이터 간의 상관관계를 악용하여 비공개 입력값을 재구성합니다. 이 프로젝트는 모델 인버전 문제를 블랙박스 환경에서의 강화학습 태스크로 공식화한 RLB-MI 공격을 재생산하는 데 중점을 둡니다.

> **원본 논문:** "Reinforcement Learning-Based Black-Box Model Inversion Attacks" (Han et al., CVPR 2023)

1.1 위협 모델 (Threat Model)

- **공격자 목표:** 분류기로부터 타겟 대상 y의 인식 가능한 얼굴 이미지를 재구성.
- **접근 수준:** 타겟 분류기에 대한 블랙박스 접근 (쿼리 기반, 출력 확률 벡터만 관측 가능).
- **보조 지식:** 동일한 데이터셋(CelebA 또는 FaceScrub)으로 훈련되었으나 타겟과 겹치지 않는 신원(identities)을 가진 사전 훈련된 생성자 G.

1.2 우리의 기여 (Our Contributions)

- **충실한 재생산:** CelebA 및 FaceScrub 데이터셋에서 SAC 에이전트를 사용한 RLB-MI 구현.
- **ArcFace 통합:** 각도 마진(angular margin)을 사용하여 견고한 얼굴 분류기를 훈련하기 위한 ArcFace 손실 함수 사용.
- **온도 스케일링 분석:** 추론 시 스케일 팩터(s=16 vs s=64)가 공격 성능에 미치는 영향 조사.

1.3 공격 예시

다음은 성공적인 모델 인버전 공격의 예시입니다. 위쪽 행은 실제 비공개 훈련 이미지(타겟 신원)를 보여주며, 아래쪽 행은 RL 에이전트가 분류기 쿼리만을 사용하여 재구성한 이미지를 보여줍니다.

타겟 (비공개 훈련 데이터)

복구됨 (공격으로 생성됨)

그림 1: 모델 인버전 공격 결과. 파란 테두리 = 실제 비공개 데이터, 빨간 테두리 = 재구성된 이미지.

2. 시스템 아키텍처 (System Architecture)

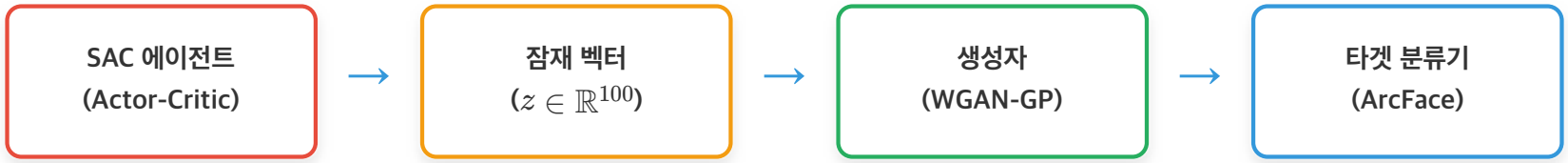

그림 1: 공격 파이프라인. SAC 에이전트는 타겟 클래스 신뢰도를 최대화하는 이미지를 생성하는 z 벡터를 찾기 위해 GAN 잠재 공간을 탐색합니다.

2.1 ArcFace 분류기

우리는 **ArcFace (Additive Angular Margin Loss)**를 사용하여 얼굴 분류기를 훈련시켰습니다. 이는 훈련 중 타겟 클래스에 각도 마진 m을 추가하여 변별력을 높입니다:

$$L = -\log \frac{e^{s(\cos(\theta_y + m))}}{e^{s(\cos(\theta_y + m))} + \sum_{j \neq y} e^{s \cos \theta_j}}$$

- **s (스케일 팩터):** 로짓의 크기를 제어 (기본값: 훈련 중 64.0)
- **m (각도 마진):** 타겟 클래스 각도에 페널티 추가 (기본값: 0.5 라디안 ≈ 28.6°)

> **핵심 통찰:** 추론 시 $s = 64$인 ArcFace는 매우 날카로운 확률 분포(신뢰도 > 99%)를 생성하여 보상 신호를 희소하게 만듭니다. 우리는 공격 중에 $s = 16$을 사용하여 보다 보정된 확률을 얻는 것을 조사했습니다.

3. 방법론 (Methodology)

3.1 MDP 공식화

- **상태 (s_t):** 현재 잠재 벡터 $z_t \in \mathbb{R}^{100}$
- **행동 (a_t):** 다음 잠재 벡터 (잠재 공간에서의 직접적인 점프)
- **전이:** $z_{t+1} = \alpha z_t + (1-\alpha)a_t$, 여기서 α는 모멘텀 팩터
- **조기 종료:** 타겟 클래스에 대한 분류기 신뢰도가 $\geq 80\%$일 때

3.2 보상 함수 (Reward Function)

보상 함수는 에이전트를 안내하기 위해 세 가지 구성 요소를 결합합니다:

$$R = w_1 r_1 + w_2 r_2 + w_3 r_3 \quad (w_1 = 2, w_2 = 2, w_3 = 8)$$

구성 요소	수식	목적
r_1 (상태 점수)	$\log P(y \mid G(z_{t+1}))$	결과 상태의 타겟 신뢰도 최대화
r_2 (행동 점수)	$\log P(y \mid G(a_t))$	높은 신뢰도의 행동 선택 장려
r_3 (구별 점수)	$\log(P(y) - \max P(\text{others}))$	타겟 클래스가 다른 클래스를 압도하도록 보장

4. 평가 지표 (Evaluation Metrics)

RLB-MI 논문을 따라 세 가지 상호 보완적인 지표를 사용하여 공격을 평가합니다:

1. 공격 정확도 (Attack Accuracy, Top-1 / Top-5)

정의: 독립적인 *평가 분류기*(타겟 분류기와 다름)가 생성된 이미지를 타겟 신원으로 올바르게 분류한 비율.

- **Top-1:** 타겟 클래스가 가장 높은 예측값일 때
- **Top-5:** 타겟 클래스가 상위 5개 예측값 안에 들 때

해석: 높을수록 좋음. 공격 성공률 및 분류기 간 전이성을 측정.

2. KNN 거리 (KNN Distance)

정의: 생성된 각 이미지와 비공개 훈련 세트의 K-최근접 이웃 간의 평균 L2 거리 (평가 분류기의 특징 공간에서 측정).

$$\text{KNN}_{\text{dist}} = \frac{1}{N} \sum \frac{1}{K} \sum \| f(\text{gen}_i) - f(\text{private}_j) \|_2$$

해석: 낮을수록 좋음. 생성된 이미지가 실제 비공개 훈련 데이터와 얼마나 유사한지 나타냄.

3. FID (Fréchet Inception Distance)

정의: InceptionV3 특징을 사용하여 생성된 이미지 분포와 비공개 이미지 분포 간의 거리를 측정.

$$\text{FID} = \|\mu_{\text{gen}} - \mu_{\text{priv}}\|^2 + \text{Tr}(\Sigma_{\text{gen}} + \Sigma_{\text{priv}} - 2(\Sigma_{\text{gen}}\Sigma_{\text{priv}})^{0.5})$$

해석: 낮을수록 좋음. 생성된 얼굴의 지각적 품질과 사실성을 측정.

5. 실험 설정 (Experimental Setup)

5.1 환경

구성 요소	사양
프레임워크	PyTorch 2.x
GPU	NVIDIA A100 (colab pro+)
언어	Python 3.10+

5.2 데이터셋

데이터셋	신원 수	이미지 수	목적
CelebA	1,000	신원당 약 30장	비공개 훈련 데이터
FaceScrub	530	신원당 약 100장	비공개 훈련 데이터

각 데이터셋은 신원이 겹치지 않게 **비공개**(분류기 훈련용)와 **공개**(GAN 훈련용)로 나뉩니다. 이는 공격자가 유사하지만 서로 다른 데이터에 접근할 수 있는 현실적인 공격 시나리오를 시뮬레이션합니다.

5.3 데이터 전처리

- **얼굴 감지 및 정렬:** MTCNN 기반 얼굴 감지 및 5개 랜드마크 정렬
- **해상도:** 모든 이미지를 64×64 픽셀로 크기 조정
- **정규화 (분류기):** Mean=[0.5177, 0.4284, 0.3803], Std=[0.3042, 0.2845, 0.2827]
- **정규화 (생성자):** 출력 범위 [-1, 1], 분류기 입력 전 [0, 1]로 변환
- **데이터 증강:** 분류기 훈련 중 무작위 수평 뒤집기 (p=0.5)

5.4 모델 구성

구성 요소	아키텍처	세부 사항
타겟 분류기	VGG16 / ResNet152 + ArcFace	$s = 64, m = 0.5$, embedding_dim=512
평가 분류기	FaceNet + ArcFace	독립적인 평가 네트워크
생성자	WGAN-GP (DCGAN 스타일)	$z_{\text{dim}} = 100$, 출력=64×64×3, 동일 데이터셋 훈련
RL 에이전트	SAC (Soft Actor-Critic)	Hidden=256×2, LR=3e-4

5.5 공격 설정

- **타겟 클래스:** 데이터셋별 가장 빈번한 상위 50개 신원
- **클래스당 에피소드:** 5,000회
- **조기 종료:** 타겟 신뢰도 $\geq 80\%$
- **모멘텀 (α):** 0.0 (직접적인 행동이 다음 상태가 됨)

- **랜덤 시드:** 재현성을 위한 고정 시드. 시간 제약으로 인해 시드 변형 실험은 수행하지 않음 (데이터셋/모델/50개 라벨당 약 4시간 소요).

5.6 분류기 정확도

다음 표는 훈련된 타겟 분류기의 각 테스트 세트에 대한 분류 정확도를 보여줍니다:

VGG16 + ArcFace

데이터셋	Top-1	Top-3	Top-5
CelebA	76.76%	84.92%	87.28%
FaceScrub	94.74%	97.37%	98.17%

ResNet-152 + ArcFace

데이터셋	Top-1	Top-3	Top-5
CelebA	62.62%	72.67%	75.63%
FaceScrub	92.00%	95.49%	96.34%

Face.evoLVe (평가 분류기)

데이터셋	Top-1	Top-3	Top-5
CelebA	65.98%	74.63%	77.40%
FaceScrub	90.86%	94.63%	95.94%

참고: 평가 분류기는 타겟 분류기와 독립적이며 공격 전이성을 측정하는 데 사용됩니다.

6. 결과 (Results)

6.1 추론 시 온도 스케일링

우리는 공격 추론 중 ArcFace 스케일 팩터의 영향을 조사했습니다:

스케일 (s)	확률 분포	보상 신호	예상 효과
s=64 (기존)	매우 날카로움 (>99%)	희소함, 이진(binary) 유사	빠른 수렴이나 지역 최적점(local optima)에 빠짐
s=16 (제안)	보정됨 (~70-90%)	유익함, 부드러움	더 나은 탐색, 더 높은 품질

6.2 주요 결과

우리는 데이터셋(CelebA, FaceScrub), 타겟 모델(VGG16, ResNet152), 스케일 팩터(s=64, s=16)의 모든 조합을 평가했습니다. **추론 시 로짓 스케일링(s=16)은 대부분의 구성에서 공격 성능을 일관되게 향상시켰습니다.** 특히, s=16을 사용했을 때 Top-1 정확도가 최대 +18%p(FaceScrub + ResNet152) 향상되었고, FID 점수가 최대 20점 감소하여 더 나은 품질의 재구성을 나타냈습니다.

데이터셋	타겟 모델	스케일	Top-1 정확도 (%)	Top-5 정확도 (%)	KNN 거리	FID
CelebA	VGG16	s=64	16.00	32.00	1.2377	99.70
CelebA	VGG16	s=16	**22.00**	**52.00**	**1.1629**	**79.82**
CelebA	ResNet152	s=64	**18.00**	26.00	1.2164	**76.99**
CelebA	ResNet152	s=16	16.00	**30.00**	**1.1965**	78.83
FaceScrub	VGG16	s=64	22.00	34.00	1.1837	91.57
FaceScrub	VGG16	s=16	**34.00**	**50.00**	**1.1017**	**76.37**
FaceScrub	ResNet152	s=64	8.00	34.00	1.2291	76.88
FaceScrub	ResNet152	s=16	**26.00**	**44.00**	**1.1488**	**67.93**

7. 논의 (Discussion)

7.1 보상 형성으로서의 온도 스케일링

추론 중 ArcFace 스케일 팩터를 s=64에서 s=16으로 줄이는 것은 **암시적인 보상 형성(implicit reward shaping)** 역할을 합니다. s=64일 때는 소프트맥스 출력이 매우 뾰족하여 잠재 공간의 작은 변화가 보상 차이를 거의 만들지 못합니다. s=16일 때는 확률 분포가 더 부드러워져 RL 에이전트에게 더 유익한 그래디언트 신호를 제공합니다.

7.2 이상 사례: 공격이 "성공"했으나 실패한 경우

우리는 생성된 이미지가 높은 신뢰도로 타겟 분류기를 속이지만, **이미지 자체는 심하게 왜곡되거나 손상되어** 사람 얼굴과 전혀 닮지 않은 흥미로운 사례를 관찰했습니다. 아래는 이러한 이상 현상의 예시입니다:

그림: 손상되거나 흐릿한 이미지가 여전히 높은 타겟 분류기 신뢰도를 달성하는 이상 사례.

분석: 이 현상은 두 가지 요인에 기인할 수 있습니다:

- **GAN 잠재 매니폴드 제한:** 모드 붕괴(mode collapse)로 인해 GAN의 잠재 공간에 특정 타겟 신원에 대한 적절한 얼굴 이미지와 매핑되는 벡터가 부족합니다. RL 에이전트는 생성자가 손상된 출력을 생성하는 영역으로 이동합니다.
- **OOD 영역에서의 분류기 과신:** 에이전트가 분포 외(OOD) 이미지를 생성하는 잠재 영역을 탐색할 때, 분류기는 낯선 영역에서 보정이 부족하여 가짜 높은 신뢰도를 보입니다. 에이전트는 본질적으로 의미 있는 이미지를 생성하지 않으면서 분류기의 결정 경계를 악용하는 "적대적" 잠재 벡터를 찾습니다.

7.3 한계점

- **생성자 용량:** 공격 품질은 GAN이 다양하고 사실적인 얼굴을 생성하는 능력에 의존합니다.

- **쿼리 효율성:** 타겟 신원당 수천 번의 쿼리가 필요합니다.
- **해상도:** 현재 구현은 64×64 이미지를 사용하며, 고해상도는 더 많은 연산을 필요로 합니다.

7.4 향후 연구

- **보상 형성 탐색:** 온도 스케일링 외에 다양한 보상 형성 기술(예: potential-based shaping, 커리큘럼 학습)을 비교하면 추가적인 성능 향상을 얻을 수 있습니다.
- **대체 RL 알고리즘:** 타겟 신원별로 별도의 훈련이 필요한 라벨별 공격 특성상 SAC만 평가할 수 있었습니다. DDPG나 TD3와 같은 다른 연속 제어 알고리즘과 비교하는 것은 계산 비용이 들지만 가치 있는 향후 연구가 될 것입니다.
- **최신 생성 모델:** GAN은 모드 붕괴에 취약하므로, 잠재 확산 모델(예: VAE 잠재 공간을 갖춘 Stable Diffusion)을 탐색하면 재구성의 다양성과 품질을 향상시킬 수 있습니다.

8. 팀원의 역할

팀원	기여 내용
김대원	• 프로젝트 주제 선정 및 문제 정의 • GAN (WGAN-GP) 훈련 및 구현 • 환경 설정 및 데이터 전처리 • ArcFace 손실 통합 분류기 훈련
최서빈	• SAC 에이전트 훈련 및 하이퍼파라미터 튜닝 • 로짓 온도 스케일링 제안 및 실험 • 보고서 작성 및 문서화

9. 결론 (Conclusion)

우리는 RLB-MI 공격 프레임워크를 성공적으로 재생산하여, 강화학습이 블랙박스 분류기를 효과적으로 공략하여 비공개 훈련 데이터를 재구성할 수 있음을 입증했습니다. 추론 시 온도 스케일링에 대한 조사를 통해, **분류기의 신뢰도 점수를 보정하는 것**이 RL 에이전트에게 더 유익한 보상 신호를 제공하여 공격 성능에 상당한 영향을 줄 수 있음을 밝혔습니다.

참고 문헌 (References)

1. Han et al., "Reinforcement Learning-Based Black-Box Model Inversion Attacks", CVPR 2023
2. Deng et al., "ArcFace: Additive Angular Margin Loss for Deep Face Recognition", CVPR 2019
3. Haarnoja et al., "Soft Actor-Critic: Off-Policy Maximum Entropy Deep RL", ICML 2018
