ViT 기반 모델 역전 공격 및 방어 기법들에 대한 연구

유미선 [1], 백윤흥 [1]
[1]서울대학교 전기정보공학부, 서울대학교 반도체 공동연구소

msyu@sor.snu.ac.kr, ypaek@snu.ac.kr

Survey of the Model Inversion Attacks and Defenses to ViT

Miseon Yu[1], Yunheung Peak[1]
[1]Dept. of Electrical and Computer Engineering and Inter-University Semiconductor Research Center(ISRC), Seoul National University

요 약

ViT(Vision Transformer)는 트랜스포머 구조에 이미지를 패치들로 나눠 한꺼번에 인풋으로 입력하는 모델이다. CNN 기반 모델보다 더 적은 훈련 계산량으로 다양한 이미지 인식 작업에서 SOTA(State-of-the-art) 성능을 보이면서 다양한 비전 작업에 ViT 를 적용하는 연구가 활발히 진행되고 있다. 하지만, ViT 모델도 AI 모델 훈련시에 생성된 그래디언트(Gradients)를 이용해 원래 사용된 훈련 데이터를 복원할 수 있는 모델 역전 공격(Model Inversion Attacks)에 안전하지 않음이 증명되고 있다. CNN 기반의 모델 역전 공격 및 방어 기법들은 많이 연구되어 왔지만, ViT 에 대한 관련 연구들은 이제 시작 단계이고, CNN 기반의 모델과 다른 특성이 있기에 공격 및 방어 기법도 새롭게 연구될 필요가 있다. 따라서, 본 연구는 ViT 모델에 특화된 모델 역전 공격 및 방어 기법들의 특징을 서술한다.

1. 서론

최근 BERT, GPT 와 같은 트랜스포머(Transformer) 기반 모델의 성능이 자연어 영역에서 주목받으면서 트랜스포머 모델을 다양한 영역으로 확장하려는 시도가 지속되고 있다. 그 중 하나가 컴퓨터 비전 분야에 트랜스포머 모델을 적용한 ViT(Vision Transformer)이다. ViT 는 트랜스포머 구조에 이미지를 패치들로 나눠 한꺼번에 인풋으로 입력하는 모델이다. 과거 컴퓨터 비전 영역에서는 주로 CNN(Convolution Neural Network) 기반의 모델이 지배적이었다. 하지만, ViT 가 더 적은 훈련 계산량으로 다양한 이미지 인식 작업에서 SOTA(State-of-the-art) 성능을 보이면서 주목받게 되었고[1], 이후로도 다양한 비전 작업에 트랜스포머를 적용하는 연구가 활발히 진행되고 있다.[2]

고도의 모델을 발전시키기 위해서는 여전히 많은 데이터가 필요하고 모델 자체가 거대해질 필요가 있다. AI 모델은 훈련과정을 통해 다양한 데이터에 대한 정보를 습득하게 된다. 즉, 이를 이용해 훈련 데이터의 정보를 유추할 수 있는 가능성이 존재한다. 그런 취약점 중 하나가 AI 모델 훈련시에 생성된 그래디언트(Gradients)를 이용해 원래 사용된 훈련 데이터를 복원할 수 있는 모델 역전 공격(Model Inversion Attacks) 혹은 그래디언트 역전 공격(Gradient Inversion Attacks)이다.[3] 해당 공격방법은 컴퓨터 비전 영역에서 CNN 기반 모델을 대상으로 주로 연구되어 왔는데, [4]를 통해 ViT 모델 또한 이 공격에 대해 정보 유출 위협이 존재한다는 것이 밝혀졌다. 모델이 거대해지면서 분산 학습(Distributed Learning) 혹은 연합 학습(Federated Learning)이 제시되면서 모델의 그래디언트 공유되는 경우가 있기 때문에 이때 그래디언트의 취약점을 잘 알고 예방하는 것이 개인 데이터 보호를 위해 필요하다.

현재, CNN 기반의 모델 역전 공격 및 방어 기법들은 많이 연구되어 왔지만, ViT 에 대한 관련 연구들은 이제 시작 단계이고, CNN 기반의 모델과 다른 특성이 있기에 공격 및 방어 기법도 새롭게 연구될 필요가 있다. 아직 ViT 에 초점을 맞춰 모델 역전 공격에 관해 정리한 논문은 없기 때문에 본 논문을 통해 해당 부분을 정리하고 최근 연구 동향을 소개하고자 한다.

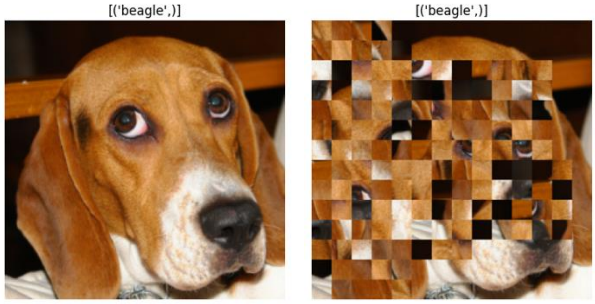
2. 비전 트랜스포머 모델 역전 공격

[4]는 ViT 기반 모델에 특화된 모델역전공격인

GradViT 를 제시하였다. 해당 공격은 원래 CNN 기반 모델에서 공격 성능이 가장 좋았던 GradInversion[5] 보다 높은 공격 성능을 ViT 모델에서 보였다. 기존 커널 기반이 아닌 어텐션(Attention) 구조를 이용해 이미지의 위치 정보를 패치(Patch)별로 기억하는 ViT 모델의 특성을 활용하였다. 이를 위해 기존 반복 기반 방법에 이미지의 위치를 적절히 배합하는 Patch Prior Regularization 을 추가적으로 넣어 원래의 이미지에 더욱 가깝게 복원할 수 있도록 하였다. 또한, 기존의 배치 정규화(Batch Normalization)를 쓰는 CNN 기반 모델과 다르게 ViT 는 레이어 정규화(Layer Normalization)를 사용하는 차이점을 고려했다. 따라서, 목표 모델을 통해 생성된 배치의 통계적 특성이 아닌 기존에 Pre-trained 된 CNN 기발 모델을 이용해 거기서 추출된 배치들의 평균, 분산을 이용해 복원 이미지가 좀 더 실제 이미지 같도록 만들었다. 이로써 CNN 기반 모델과 다른 ViT 구조에 특화된 새로운 방식의 모델역전공격방법을 제안했다. [6]은 재귀 기반 데이터 복구 방식을 주로 이용한 것으로 각 레이어마다의 인풋 피처(input feature)를 재귀적으로 계산하는 방법을 제안했다. 해당 논문은 셀프-어텐션(self-attention)의 아웃풋과 로스(loss)값을 알면 해당 레이어의 인풋 피처를 수학적으로 계산할 수 있음을 증명하였다. 이를 통해 포지션 임베딩(position embedding)의 아웃풋까지 재귀적으로 계산하고 이를 이용해 원래의 데이터를 복구하는 방법을 제안했다.

3. 방어 기법

[1] 은 ViT 의 어텐션 메커니즘이 기존 CNN 베이스의 모델보다 모델 역전 공격에 대해 더 취약함을 실험적으로 주장하였다. 본 저자는 실험을 통해 MLP(Multi Layer Perceptron) 영역만 사용했을 때보다 MSA(Multi Self-Attention) 영역만 공격에 사용했을 때 이미지 복구 공격에 더 취약함을 주장한다. 따라서, ViT 가 가지고 있는 어텐션 영역이 해당 위협에 더 취약함을 나타냈다. 그래디언트의 어텐션 영역 혹은 전체 그래디언트에 동형 암호 기법을 써서 가리는 방어 기법이 연구될 수 있다. 혹은 그래디언트의 통계적 특성은 유지시키면서 일정량의 노이즈를 추가하는 차분 프라이버시 방어기법도 적용될 수 있다. [7]은 ViT 가 이미지의 전체적 특성을 한번에 파악하는 어텐션 메커니즘에 초점을 맞춰 패치 분할과 포지션 임베딩의 특성을 이용한 방어기법을 제시했다. ViT 에서 포지션 임베딩은 인풋 패치들의 이차원 공간 정보를 학습한다. 이 인풋 패치들의 순서를 랜덤하게 일정량 변화시키는 랜덤 셔플(random shuffle) 기법을 적용한다. 변화된 패치들에는 unknown 포지션 인코딩의 값을 할당했다. 따라서, 공격자가 패치들이 뒤섞인 위치를 모르기 때문에 원래의 이미지를 완벽히 복원할 수 없게 된다. 대신, 변화된 패치 순서가 모델 학습에 영향을 주지 않도록 위치가 변화되지 않은 패치에는 원래의 포지션 인코딩 값을 넣어주었다. 또한, 위치가 변화되지 않은 패치의 임베딩 값에서 원래의 위치를 복원하는 레이어를 추가해 해당 레이어가 복원하는 위치와 원래 위치와의 차이를 나타내는 DAL(dense absolute localization) loss 를 추가하여 모델 훈련 정확도가 심하게 하락하지 않도록 보완하였다. 그림 1 은 [7]에서 적용한 랜덤 셔플을 실제 이미지에 적용한 모습이다. 그림 1 의 오른쪽 이미지는 이미지 전체의 70%가 랜덤하게 섞인 모습을 볼 수 있다.



(그림 1) 왼쪽 : 원본 이미지, 오른쪽 : 70%의 비율로 셔플된 이미지

4. 결론

본 논문은 ViT 모델에 특화된 모델 역전 공격 및 방어 기법들을 서술하였다. 기존 CNN 기반 모델과 달리 어텐션 구조와 포지션 임베딩 레이어를 가지고 있기 때문에, 모델 역전 공격 및 방어 기법들도 그에 맞게 조정될 수 있음을 확인했다. 이를 통해 ViT 기반 모델을 사용해도 여전히 공유된 그래디언트에 개인 정보 탈취 위협이 존재함을 알 수 있다. 안전한 머신러닝 산업을 위해 해당 공격에 대해 알고 이에 따른 대응방법을 연구해야 한다.

ACKNOWLEDGEMENT

이 논문은 2023 년도 BK21 FOUR 정보기술 미래인재 교육연구단에 의하여 지원되었음. 본 연구는 반도체 공동연구소 지원의 결과물임을 밝힙니다. 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학 ICT 연구센터육성지원사업의 연구결과로 수행되었음 (IITP-2023-2020-0-01602). 이 논문은 2023 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2022-0-00516, 국가통계데이터에 적용 가능한 차등정보보호 개념을 도출하고 통계분석의 유용성을 보장해야 하는 문제 해결)

참고문헌

[1] Dosovitskiy, Alexey et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.” ArXiv abs/2010.11929 (2020).

[2] K. Han et al., "A Survey on Vision Transformer," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 1, pp. 87-110, 1 Jan. 2023.

[3] 유미선, 이영한, 한우림, 백윤흥. “모델 역전 공격 및 방어 기법들에 대한 연구”, 정보보호학회, 2022.11, 4.

[4] Hatamizadeh, Ali, et al. "Gradvit: Gradient inversion of vision transformers." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[5] Yin, Hongxu, et al. "See through gradients: Image batch recovery via gradinversion." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

[6] Lu, Jiahao, et al. "APRIL: Finding the Achilles' Heel on Privacy for Vision Transformers." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[7] Ren, Bin, et al. "Masked Jigsaw Puzzle: A Versatile Position Embedding for Vision Transformers." arXiv e-prints (2022): arXiv-2205.