

Textual Inversion을 활용한 Adversarial Prompt 생성 기반 Text-to-Image 모델에 대한 멤버십 추론 공격*

오윤주,^{1*} 박소희,¹ 최대선^{2*}
^{1,2}숭실대학교 (대학원생, 교수)

Membership Inference Attack against Text-to-Image Model Based on Generating Adversarial Prompt Using Textual Inversion*

Yoonju Oh,^{1*} Sohee Park,¹ Daeseon Choi^{2*}
^{1,2}Soongsil University (Graduate student, Professor)

요약

최근 생성 모델이 발전함에 따라 생성 모델을 위협하는 연구도 활발히 진행되고 있다. 본 논문은 Text-to-Image 모델에 대한 멤버십 추론 공격을 위한 새로운 제안 방법을 소개한다. 기존의 Text-to-Image 모델에 대한 멤버십 추론 공격은 쿼리 이미지의 caption으로 단일 이미지를 생성하여 멤버십을 추론하였다. 반면, 본 논문은 Textual Inversion을 통해 쿼리 이미지에 personalization된 임베딩을 사용하고, Adversarial Prompt 생성 방법으로 여러 장의 이미지를 효과적으로 생성하는 멤버십 추론 공격을 제안한다. 또한, Text-to-Image 모델 중 주목받고 있는 Stable Diffusion 모델에 대한 멤버십 추론 공격을 최초로 진행하였으며, 최대 1.00의 Accuracy를 달성한다.

ABSTRACT

In recent years, as generative models have developed, research that threatens them has also been actively conducted. We propose a new membership inference attack against text-to-image model. Existing membership inference attacks on Text-to-Image models produced a single image as captions of query images. On the other hand, this paper uses personalized embedding in query images through Textual Inversion. And we propose a membership inference attack that effectively generates multiple images as a method of generating Adversarial Prompt. In addition, the membership inference attack is tested for the first time on the Stable Diffusion model, which is attracting attention among the Text-to-Image models, and achieve an accuracy of up to 1.00.

Keywords: Generative Model, Text-to-Image Model, Membership Inference Attack, Adversarial Prompt, Textual Inversion

Received(10. 24. 2023), Modified(11. 28. 2023),
Accepted(11. 28. 2023)

* 본 논문은 이 논문은 2023년도 정부(과학기술정보통신부)의
재원으로 정보통신기획평가원의 지원(No. 2021-0-00511,
엣지 AI 보안을 위한 Robust AI 및 분산 공격탐지기술 개

발)과 2023년도 정부(과학기술정보통신부)의 재원으로 한국
연구재단의 지원을 받아 수행된 연구임 (No. 2020R1A2C1
014813)

† 주저자, ohyoonju@soongsil.ac.kr

‡ 교신저자, sunchoi@ssu.ac.kr(Corresponding author)

1. 서 론

최근 생성 모델(Generative Model)이 많이 발전되면서 동시에 생성 모델에 대한 보안 중요성이 강조되고 있다. 생성 모델을 위협하는 공격으로 백도어(Backdoor) 공격, 모델 추출(Model Extraction) 공격, 멤버십 추론(Membership Inference) 공격 등이 있다. 그중 멤버십 추론 공격은 공격자에게 주어진 쿼리 데이터가 member인지 non member인지에 대한 멤버십 여부를 추론하는 공격이다. member는 target 모델의 훈련에 사용한 훈련 dataset의 일부 데이터를 말하며, non member는 훈련에 사용하지 않은 데이터를 말한다. 훈련 데이터 정보를 알아낼 수 있는 멤버십 추론 공격은 심각한 프라이버시 침해를 일으킬 수 있으며, 이를 이용하여 다른 공격에 악용할 위험이 있다. 따라서 점점 발전하고 있는 생성 모델에서의 멤버십 추론 공격은 생성 모델이 이에 대한 방어 방법을 모색·구현하는 연구를 진행하기 위해 활용될 수 있으며, 결과적으로 생성 모델의 보안성을 강화할 수 있다.

Wu 등[1]은 Text-to-Image(T2I) 모델에 대한 멤버십 추론 공격을 최초로 제안하였으며, 현재까지 유일하게 소개된 T2I 모델에 대한 멤버십 추론 공격이다. 쿼리 이미지의 caption을 T2I 모델의 입력 prompt로 사용하여 단일 생성 이미지를 생성하였으며, 멤버십 추론 공격의 직관(Intuition)으로 생성 이미지에 대한 3가지 성능을 이용하여 4가지 공격을 제안하였다. 생성 이미지의 품질, 쿼리 이미지와 생성 이미지 간 거리, caption과 생성 이미지 간 거리를 이용하였으며, diffusion 기반의 T2I 모델인 LDM[2]과 sequence to sequence 기반 T2I 모델인 DALL-E mini[3]에 대해 실험을 진행하였다.

하지만 단일 생성 이미지를 사용하여 멤버십을 추론하면 멤버십을 판단할 근거가 하나의 이미지뿐이므로, 여러 장의 생성 이미지를 사용하는 경우보다 멤버십을 잘못 예측할 확률이 크다. 또한, T2I 모델은 입력한 prompt를 기반으로 이미지를 생성하는데 captioning 모델을 통해 생성한 caption이 구체적인 경우와 그렇지 않은 경우 사이에 생성 이미지의 성능 차이가 발생한다. 다시 말해, non member의 구체적인 caption이 member의 구체적이지 않은 caption보다 더 좋은 성능을 갖는 이미지를 생성하게 되어 멤버십을 잘못 예측할 확률이 크다는 것이다. 이는 멤버십 외 다른 요인으로 생기는 동작 차이이므

로 멤버십 추론 공격에 좋지 않은 영향을 줄 수 있다.

따라서 본 논문은 'Textual Inversion을 활용한 Adversarial Prompt 생성 기반 T2I 모델에 대한 멤버십 추론 공격'을 제안한다. Textual Inversion을 활용한 쿼리 이미지의 토큰 임베딩을 T2I 모델의 초기 입력으로 사용하며, Adversarial Prompt 생성 방법을 이용하여 T2I 모델의 입력을 최적화한다. Adversarial Prompt는 prompt를 입력으로 사용하는 모델의 target 동작을 유도하는 prompt를 말하며, 이 prompt를 생성하기 위해 objective function을 정의하고 prompt를 찾는 과정을 이용하여 여러 장의 생성 이미지를 생성하고자 한다. Textual Inversion은 T2I 모델에 대한 personalization(개인화) 방법으로, 쿼리 이미지 내 개체 개념을 학습하여 토큰 임베딩 $S^*:e^*$ 으로 표현한다. 토큰 임베딩이 포함된 prompt 'a photo of a S^* '를 T2I 모델의 입력으로 사용하면, 학습한 쿼리 이미지의 개념을 나타내는 생성 이미지를 출력할 수 있다. Textual Inversion을 T2I 모델의 입력으로 사용하면, 모두 동일한 prompt 'a photo of a S^* '를 사용하면서 쿼리 이미지를 나타내는 이미지를 생성할 수 있으므로 Wu 등[1]의 연구에서 쿼리 이미지의 caption을 입력으로 사용했을 때 발생할 수 있는 문제점을 보완할 수 있다.

정리하면, 본 논문의 연구 목적과 멤버십 추론 공격의 직관은 Wu 등[1]의 연구와 동일하지만, T2I 모델의 입력 유형과 생성 이미지를 출력하는 방법에 대해 새롭게 접근하는 멤버십 추론 공격을 제안하고자 한다. 이에 따른 본 논문의 기여는 다음과 같다.

- 최근 생성 모델로 주목받고 있는 Stable Diffusion 모델에 대해 멤버십 추론 공격을 최초로 수행한다.
- 기존 연구[1]의 T2I 모델의 입력 생성 방법에 대해 문제점을 제기하고, Textual Inversion을 이용하여 보완하는 방법을 제시한다.
- 여러 개의 이미지를 효과적으로 생성하기 위해 Adversarial Prompt 생성 방법을 사용하고, Textual Inversion과 함께 적용한 멤버십 추론 공격을 제안하여 높은 성능을 달성한다.

본 논문의 구성은 다음과 같다. 2장에서는 연구 배경 및 관련 연구로 멤버십 추론 공격과 T2I 모델에 대한 멤버십 추론 공격에 대한 기존 연구, Adversarial Prompt 생성 방법과 이를 이용한

T2I 모델의 입력 최적화 방법, Textual Inversion에 대해 설명한다. 3장에서는 연구 동기와 제안 방법에 대해 설명한다. 4장에서는 이에 대한 실험 환경 설정과 threshold 설정 방법 그리고 실험 결과를 서술하며, 마지막으로 5장의 결론으로 논문을 맺는다.

II. 배경 및 관련 연구

2.1 Text-to-Image 모델에 대한 멤버십 추론 공격

2.1.1 멤버십 추론 공격 (MIA, Membership Inference Attack)

멤버십 추론 공격은 해당 데이터가 target 모델의 훈련에 사용된 member인지 훈련에 사용되지 않은 non member인지에 대한 여부를 추론하는 공격이다. 멤버십 추론 공격의 기본 아이디어는 target 모델에 대해 member를 쿼리했을 때와 non member를 쿼리했을 때의 동작 차이이다. target 모델에 따라 동작 차이는 다르게 구성되며, 예를 들어 target 모델이 분류기인 경우 member의 confidence score가 non member보다 높을 것이라는 동작 차이를 이용하여 멤버십을 추론할 수 있다.

Hayes 등[4]은 생성 모델에 대한 최초의 멤버십 추론 공격을 제안하였다. GAN(Generative Adversarial Networks)에 대한 멤버십 추론 공격으로 target 모델의 Discriminator를 이용하는 white-box 방법, 보조 지식으로 훈련한 Discriminator를 이용하는 방법, 보조 지식이 없지만 쿼리 이미지와 노이즈로 훈련한 Discriminator를 이용하는 black-box 방법을 제시한다. Liu 등[5]은 쿼리 이미지와 생성 이미지 간 복원 손실을 이용한 GAN과 VAE(Variational AutoEncoder)의 멤버십 추론 공격을 제안하였다. 이 외에도 생성 모델 GAN, VAE에 대한 멤버십 추론 공격을 다양한 방식으로 제안한 연구들[6,7]이 있다.

2.1.2 Text-to-Image 모델에 대한 멤버십 추론 공격

생성 모델 GAN, VAE에 대한 멤버십 추론 공격 연구가 활발하게 진행된 반면, Text-to-Image에 대한 멤버십 추론 공격을 다룬 이전 연구는 Wu 등[1]의 연구가 유일하다. 이 연구는 생성 모델 중에

서 LDM과 DALL-E mini에서 실험을 진행하였으며, T2I 모델의 출력(생성 이미지)만으로 멤버십 추론 공격을 진행하였다. 구체적인 공격 과정은 다음과 같다.

쿼리 이미지가 주어지면 Blip 또는 ClipCap과 같은 captioning 모델을 이용하여 caption을 생성하고, 생성한 caption을 target T2I 모델에 입력하여 단일 이미지를 생성한다. 그런 다음, 3가지 직관을 이용한 4가지 공격에 따라 구한 동작 차이를 공격 모델에 입력하여 멤버십을 추론한다. 이때, 공격 모델은 보조 지식으로 미리 훈련시키며, Wu 등[1]이 제안한 4가지 공격은 다음과 같다.

- (1) 생성 이미지 품질은 member가 non member보다 높다는 직관을 이용하여, 생성 이미지를 입력하거나 생성 이미지의 임베딩을 입력한다.
- (2) 쿼리 이미지와 생성 이미지 사이의 복원 오차는 member가 non member보다 작다는 직관을 이용하여, 쿼리 이미지와 생성 이미지 사이의 거리를 입력하거나 쿼리 이미지의 임베딩과 생성 이미지의 임베딩 사이의 거리를 입력한다.
- (3) 생성 이미지의 입력 prompt 충실도는 member가 non member보다 높다는 직관을 이용하여, 쿼리 이미지 caption의 임베딩과 생성 이미지의 임베딩 사이의 거리를 입력한다.
- (4) (1)~(3) 중에서 임베딩 입력을 사용하는 경우를 모두 입력한다. 이는 실험적으로 이미지를 사용하는 pixel 수준의 입력보다 임베딩을 사용하는 semantic 수준의 입력을 사용했을 때의 공격 성능이 더 좋은 것을 발견하였기 때문이다.

2.2 Adversarial Prompt 생성 연구를 이용한 Text-to-Image의 입력 최적화 방법

2.2.1 Adversarial Prompt 생성 연구

prompt를 입력으로 사용하는 모델에서 target 동작을 유도하는 prompt를 'Adversarial Prompt'라고 한다. T2I 모델에서는 target 이미지를 생성하는 prompt를 Adversarial Prompt라고 할 수 있으며, 특정 이미지가 생성되도록 정의한 objective function을 최적화하여 prompt를 생성한다. 최근 들어 T2I 모델의 Adversarial Prompt에 대한 연구[8-11]가 많이 소개되고 있으며, 본 논문은 Adversarial Prompt의 생성 방법

을 활용하여 T2I 모델의 입력을 최적화한다.

Maus 등[8]은 black-box 환경의 T2I 모델에서 Adversarial Prompt를 찾는 프레임워크를 제시하였다. target 동작은 prompt에서 설명하는 class와 다른 class에 속하는 이미지를 생성하는 것이다. 예를 들어, 'mountain'을 생성하는 prompt 'picture of a mountain'을 약간 변형한 prompt 'turbo lhaff✓ picture of a mountain'를 T2I 모델에 입력하면 'mountain'이 아닌 'dog' 이미지가 생성된다. 이는 생성 이미지의 분류 결과가 target class에 가까워지도록 objective function을 정의하여 prompt를 최적화한 것이다. 이 연구는 모델 쿼리 기반 adversarial 공격이며, 4개의 단어를 사용하여 성공적인 Adversarial Prompt를 찾기 위해 10,000개의 쿼리를 요구한다.

이에 반해, Zhuang 등[9]은 쿼리가 없는 공격에 초점을 맞추며, 5자의 perturbation만으로 성공적인 Adversarial Prompt를 찾는 방법을 제안하였다. target 동작은 prompt에 설명되어 있는 content가 수정/제거된 이미지를 생성하는 것이다. 예를 들어, 'A snake and a young man'에 5자의 perturbation이 추가된 prompt 'A snake and a young man-08=*'를 T2I 모델에 입력하면 'a young man'이 제거되어 'snake'만 존재하는 이미지가 생성된다. 이는 prompt(x)의 임베딩과 perturbation이 추가된 prompt(x')의 임베딩이 가까워지도록 objective function을 정의하여 prompt를 최적화한 것이다. prompt 최적화 방법으로는 PGD 공격, Greedy Search, Genetic Algorithm을 사용하였다.

Liu 등[10]은 diffusion 기반 T2I 모델의 실패 사례를 찾는 T2I 모델에 대한 adversarial 공격을 제안하였다. 발견한 실패 사례 중 하나는 Adversarial Prompt로, target 동작은 텍스트 CLIP score를 크게 변경하지 않고도 입력 prompt와 다른 개체의 이미지를 생성하는 것이다. 예를 들어, 'A photo of a cat [x]'를 T2I 모델에 입력하면 'cat'이 아닌 'car' 이미지를 생성하는 Adversarial Prompt를 찾을 수 있다. 이때, 'A photo of a cat [x]'과의 텍스트 CLIP score는 'A photo of a cat'과 0.742, 'A photo of a car'와 0.546으로 차이가 크지 않다. 이는 prompt의 핵심 개체인 'cat'과의 CLIP score는 최소화하고 target 개체인 'car'와의 CLIP score는 최대화하

도록 objective function을 정의하여 prompt를 최적화한 것이다.

Liu 등[11]은 이전 Adversarial Prompt 연구의 'Apoploe vesreaitais'처럼 자연스럽지 않고 더 무니 없거나, 'happeerful'('happy'+ 'cheerful')처럼 쉽게 유추가 가능한 Adversarial Prompt가 아닌 신뢰할 수 있으며 감지할 수 없는 Adversarial Prompt를 생성하는 방법을 제안하였다. target 동작은 prompt에서 설명하는 것과 상관없는 target 이미지와 유사한 이미지를 생성하는 것이다. 예를 들어, 'a hoesse is standing on gra ss'를 T2I 모델에 입력하면 target 이미지('호수 배경의 배' 이미지)와 유사한 이미지가 생성된다. 이는 target 이미지 x (호수 배경의 배' 이미지), target 이미지와 pair한 텍스트 y (a boat floating on top of a lake'), original 텍스트 z (a horse is standing on grass')가 있을 때, y 와 z 가 멀어지고 x 와 z 로 생성한 이미지가 가까워지는 것을 목표로 한다. 따라서 y 의 임베딩과 z 의 임베딩 사이의 코사인 유사도와 x 의 임베딩과 z 로 생성한 이미지의 임베딩 사이의 코사인 유사도가 커지도록 objective function을 정의하여 prompt를 최적화한다. prompt 최적화 방법으로는 Genetic Algorithm을 사용하였다.

2.2.2 Adversarial Prompt 생성 연구를 이용한 Text-to-Image의 입력 최적화 방법

2.2.1에서 소개한 연구[8-11]는 T2I 모델에서의 target 동작을 수행하기 위해 objective function을 정의하고, T2I 모델의 입력인 prompt를 최적화하여 Adversarial Prompt를 찾는 방법을 제안한다.

이를 본 논문에 적용하면, 여러 장의 생성 이미지를 효과적으로 생성하기 위해 쿼리 이미지와 유사한 이미지를 생성하는 것을 target 동작으로 설정한다. 이는 기존 연구[1]에서 생성 이미지와 쿼리 이미지의 거리가 가까울수록 member라고 예측하여, 높은 공격 성공률을 보였기 때문이다. 따라서 쿼리 이미지의 임베딩과 생성 이미지의 임베딩이 가까워지도록 objective function을 정의하여 prompt를 최적화한다. 최적화 방법으로는 일부 Adversarial Prompt 연구[9,11]에서 사용하는 Genetic Algorithm으로 prompt를 최적화하며, Genetic Algorithm은 다음 과정을 통해 원하는 솔루션을 얻는다.

1. p개의 초기 솔루션을 생성한다.
2. objective function으로 각 솔루션을 평가하여, 설정한 기준에 따라 p개 중 일부 솔루션을 선택한다.
3. 선택한 솔루션을 2개씩 짝지은 뒤, 각각 crossover하여 새로운 솔루션을 생성한다.
4. 생성한 솔루션에서 비율 m만큼 mutation을 넣어, 새로운 솔루션을 생성한다.
5. 솔루션을 업데이트하고, 과정 2~5를 반복한다.

따라서 본 논문은 Genetic Algorithm을 통해 쿼리 이미지와 생성 이미지가 가까워지도록 정의한 objective function을 최적화하여, T2I 모델의 입력을 업데이트하면서 이미지를 생성할 것이다. 제안 방법 과정에 대한 자세한 설명은 3.2에서 확인할 수 있다.

2.3 Textual Inversion

'Textual Inversion[12]'은 소수의 이미지에 포함된 개념을 캡처하는 기술로, T2I 모델에 대한 효율적인 personalization(개인화)을 가능하게 한다. Textual Inversion은 사용자가 표현하고자 하는 개념(개체 또는 스타일)이 포함된 3-5개의 이미지를 사용하여, T2I 모델의 텍스트 임베딩 공간에서 구체적인 개념을 나타내는 새로운 임베딩을 찾는 것을 목표로 한다. 사용자가 제공한 개념을 잘 표현할 수 있는 임베딩을 찾으면, 그 개념에 새로운 장면을 합성하여 원하는 이미지를 생성할 수 있다. 예를 들어, '정교한 해골이 그려진 컵' 이미지를 넣어 Textual Inversion을 진행하여, 해당 개념을 학습한 토큰 임베딩 S^* 을 얻을 수 있다. 그런 다음 prompt 'A photo of S^* on the beach'를 T2I 모델에 입력하면 바다 배경에 해골 컵이 정교한 디자인 그대로 합성된 이미지가 생성된다. 또한, 특정 스타일이 담긴 이미지에 대한 Textual Inversion을 진행하여 '□ in the style of S^* '를 T2I 모델에 입력하여 해당 스타일이 포함된 '□' 이미지를 생성할 수 있다. 개체와 스타일에 대한 Textual Inversion 임베딩을 동시에 반영하는 이미지를 생성하는 것도 가능하다.

Textual Inversion의 목표는 '생성을 안내할 수 있는 의사(pseudo) 단어 찾기'이므로, 시각적 복원에 대한 objective function을 이용한다. Diffusion 기반 T2I 모델에서 이미지를 생성할 때 사용하는 reconstruction loss를 활용하여 objective function을 정의하고, Textual Inversion의 토큰

임베딩을 찾는다. 이때, reconstruction loss는 수식 (1)을 따르고, reconstruction loss를 활용한 Textual Inversion의 objective function은 수식 (2)를 따른다. 수식을 보면, 문자열 y에 대해서만 계산하던 reconstruction loss를 Textual Inversion의 objective function에서는 문자열 y에 target 토큰 임베딩 v를 포함한 채로 reconstruction loss를 계산한다.

$$\begin{aligned} \mathcal{L}_{LDM} := & \quad (1) \\ E_{z \sim \epsilon(x), y, \epsilon \sim \mathcal{N}(0,1), t} [\|\epsilon - \epsilon_\theta(z_t, t, c_\theta(y))\|_2^2] \\ (t: \text{timestep}, z_t: \text{latent noised to time } t, \\ \epsilon_\theta: \text{denoising network}, c_\theta: \text{text encoder}, \\ y: \text{text prompt}) \end{aligned}$$

$$\begin{aligned} v^* = \underset{v}{\operatorname{argmin}} & \quad (2) \\ E_{z \sim \epsilon(x), y, \epsilon \sim \mathcal{N}(0,1), t} [\|\epsilon - \epsilon_\theta(z_t, t, c_\theta(y, v))\|_2^2] \\ (t: \text{timestep}, z_t: \text{latent noised to time } t, \\ \epsilon_\theta: \text{denoising network}, c_\theta: \text{text encoder}, \\ y: \text{text prompt}, v: \text{target embedding}) \end{aligned}$$

Gal 등[12]이 제안한 Textual Inversion이 소개된 이후, Textual Inversion에 대한 다양한 연구[13-15]가 진행되었다.

Fei 등[13]은 모델의 back-propagation(역전파) 없이 Textual Inversion을 최적화하는 방법을 제안하였다. gradient 없이 (gradient-free) 최적의 의사 토큰 임베딩을 결정하게 되면 고차원의 임베딩을 다루기 힘들다는 문제가 생긴다. Fei 등[13]은 사전 훈련된 CLIP을 이용해서 초기 임베딩을 생성하고 gradient-free 최적화를 위해 널리 사용되는 CMA-ES(Covariance Matrix Adaptation Evolution Strategy)를 이용하여 임베딩을 업데이트하여, gradient-free Textual Inversion이 가능하다는 것을 보였다.

Gal 등[14]은 Textual Inversion보다 빠르고 간단하게 personalization 할 수 있는 domain tuning 접근법을 제안하였다. 개념의 세부 사항을 학습하는 것이 아니라, 개념과 가까운 domain의 큰 개념에 대한 가중치 set를 학습한다. 예를 들어, 고양이 개념을 가진 큰 dataset LSUN-Cat에서 모델을 조정하여, unseen 고양이 개념에 더 쉽게 personalization 할 수 있도록 한다. 이 접근법을

이용하면, 단일 이미지와 최소 5번의 훈련 단계만으로 personalization이 가능하다.

Melas-Kyriazi 등[15]은 2D diffusion 이미지 생성기를 이용하여 단일 이미지를 360° photographic 3D 복원으로 추출할 수 있는 방법을 제안하였다. 이때 Textual Inversion을 활용한 토큰 임베딩을 prompt로 사용하여 diffusion 모델을 조정하는 것이 제안 방법의 핵심 구성 요소라고 언급하였다.

논문은 생성 이미지와 쿼리 이미지의 유사도가 높을수록 member일 확률이 높다는 직관을 이용하여 멤버십 추론 공격을 진행한다. 따라서 쿼리 이미지에 대해 Textual Inversion한 토큰 임베딩을 T2I 모델의 초기 입력으로 사용하여, 쿼리 이미지와 유사한 생성 이미지를 얻고자 한다. Textual Inversion을 활용한 토큰 임베딩은 'a photo of a S*'라는 고정 prompt로 T2I 모델에 입력된다.

III. Textual Inversion을 활용한 Adversarial Prompt 생성 기반 Text-to-Image 모델에 대한 멤버십 추론 공격

3.1 동기

본 논문은 T2I 모델에 대한 멤버십 추론 공격의 성능을 높이기 위해 (1) 여러 개의 생성 이미지를 사용하며 Adversarial Prompt 생성 방법을 활용하여 이미지를 생성하고, (2) Textual Inversion을 활용하여 T2I 모델의 입력을 초기화한다.

기존의 유일한 T2I 모델에 대한 멤버십 추론 공격[1]은 T2I 모델의 생성 이미지만을 활용하여 높은 공격 성공률을 보였다. 이를 통해 생성 이미지가 멤버십 추론 공격 성능에 미치는 영향이 큰 것을 알 수 있다. 본 논문은 단일 생성 이미지를 이용하여 멤버십 추론 공격을 진행했던 기존 연구[1]와 달리, 여러 장의 생성 이미지를 사용하여 공격 성능을 높이고자 한다. 이는 Stable Diffusion, Imagen과 같은 Diffusion 모델이 훈련 데이터를 기억한다는 것을 밝힌 Carlini 등[16]의 연구의 영향을 받았다. Carlini 등[16]은 훈련 데이터를 활용하여 생성한 이미지에서 훈련에 사용된 이미지와 동일한 이미지(memorization이라고 명칭하는 기억된 이미지)를 찾아내었다. 이를 통해 생성 모델의 데이터 추론에 대한 위험성을 강조하였으며, 훈련 데이터를 활용하

였으므로 멤버십 추론 공격이 아니다. 구체적으로, 각 (이미지, caption) 쌍에 대해 500장의 이미지를 생성하고 생성 이미지 간 유사도를 비교하여 memorization을 찾아내었다. 단 몇 장이 아닌 수백 장의 이미지를 생성하여 memorization을 찾는 것은 '생성 이미지가 많을수록 member를 찾을 확률이 크다'는 것을 간접적으로 얘기한다. 따라서 본 논문은 단일 생성 이미지가 아닌 여러 장의 생성 이미지를 이용하여 멤버십을 추론한다.



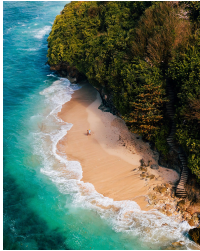



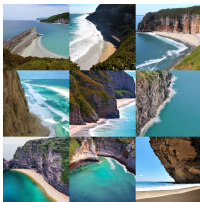


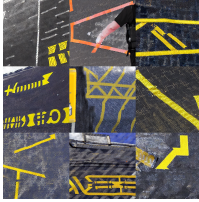
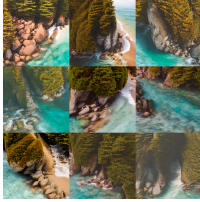

그리고 기존의 T2I 모델에 대한 멤버십 추론 공격[1]은 쿼리 이미지(또는 임베딩)와 생성 이미지(또는 임베딩)의 거리가 짧을수록 또는 쿼리 이미지 caption의 임베딩과 생성 이미지의 임베딩 사이의 거리가 짧을수록 member일 확률이 높다는 것을 입증했다. 이는 쿼리 이미지와 유사한 생성 이미지일수록 또는 T2I 모델의 입력과 유사한 생성 이미지일수록 member일 확률이 높다는 것을 의미한다.

따라서 본 논문은 쿼리 이미지의 임베딩과 생성 이미지의 임베딩 사이의 코사인 유사도를 objective function으로 정의하여 T2I 모델의 입력을 최적화하고자 한다. 이는 target 목적에 맞는 objective function을 정의하여 T2I 모델의 입력을 최적화하는 Adversarial Prompt 생성 방법을 활용한 것이며, Genetic Algorithm을 이용하여 최적화를 진행하였다.

또한, 본 논문은 T2I 모델의 입력으로 쿼리 이미지의 caption이 아닌 Textual Inversion을 통해 얻은 토큰 임베딩을 사용하고자 한다. Table 1.은 member와 non member에 대한 쿼리 이미지가 주어질 때, T2I 모델의 입력에 따른 생성 이미지를 보여준다. T2I 모델의 입력은 captioning 모델 Blip2로 생성한 쿼리 이미지의 caption을 사용하는 경우와 쿼리 이미지를 Textual Inversion하여 얻은 토큰 임베딩 S*을 활용한 prompt 'a photo of a S*'를 사용하는 경우를 비교하였다. caption을 사용한 생성 이미지의 경우, member (a), (b)의 생성 이미지보다 non member (c)의 생성 이미지가 쿼리 이미지를 더 잘 반영한 것을 볼 수 있다. member의 생성 이미지보다 non member의 생성 이미지가 쿼리 이미지와 더 유사한 것은 멤버십 외 요소인 caption이 멤버십 추론 공격 성능에 영향을 미칠 수 있음을 의미한다.

반면, Textual Inversion하여 얻은 토큰 임베딩을 사용한 생성 이미지의 경우, (a)에서 차의 앞모습

Table 1. Generated Images according to input of Text-to-Image Model

		member		non member	
		(a)	(b)	(c)	(d)
generated img	query img				
	caption of the query img				
	token embedding by Textual Inversion				

에 대한 이미지를 생성하거나 (d)에서 애니 스타일을 반영한 이미지를 생성하는 것과 같이 caption을 사용하여 생성한 이미지보다 쿼리 이미지와 유사하다. 따라서 Textual Inversion을 T2I 모델의 초기 입력으로 사용하면, 멤버십 외 요소인 caption이 생성 이미지에 미치는 영향을 줄일 수 있다.

따라서 본 논문은 Textual Inversion을 활용한 Adversarial Prompt 생성 기반 T2I 모델에 대한 멤버십 추론 공격을 제안한다.

3.2 제안 방법

본 논문은 Textual Inversion을 활용한 Adversarial Prompt 생성 기반 T2I 모델에 대한 멤버십 추론 공격을 제안한다. Adversarial Prompt 생성 기반 최적화를 위해 Genetic Algorithm을 사용하며, 본 논문의 공격 상황에서 Genetic Algorithm의 솔루션은 'T2I 모델의 입력 (토큰 임베딩)'이고, objective function은 '쿼리 이미지의 임베딩과 생성 이미지의 임베딩 간 코사인 유사도'다. 이때, 쿼리 이미지에 대해 Textual

Inversion을 진행하여 얻은 토큰 임베딩을 초기 솔루션으로 사용하는데, Textual Inversion 토큰 임베딩은 학습을 통해 1개만 얻을 수 있으며 T2I 모델은 1개의 입력만으로 여러 개의 이미지를 생성할 수 있다. 따라서 p개의 초기 솔루션으로 시작하는 일반적인 Genetic Algorithm과 달리 본 논문은 1개의 초기 솔루션을 사용한다. 그러므로 첫 번째 iteration에서는 objective function에 대한 평가가 없으며, 다음 iteration에서 사용할 토큰 임베딩 후보를 생성한다.

본 논문이 제안하는 Textual Inversion을 활용한 Adversarial Prompt 생성 기반 T2I 모델에 대한 멤버십 추론 공격을 위한 준비(이미지 생성) 과정은 다음과 같다.

1. 쿼리 이미지에 대한 Textual Inversion을 진행하여 1개의 초기 토큰 임베딩 v^* 를 생성한다.
2. 토큰 임베딩을 설정한 crossover 값 c 만큼 복제하여, 2개씩 짝지은 뒤 c 번의 crossover를 진행한다. 매번 4가지 방법 one point, multi point, uniform, arithmetic 중 하나를 랜덤

- 하게 선택하여 crossover를 진행하며, 각 방법에 대한 설명은 다음과 같다.
- one point : 한 개의 인덱스를 기준으로 앞, 뒤 값을 서로 교차한다.
 - multi point : 여러 개의 인덱스를 기준으로 앞, 뒤 값을 서로 교차한다.
 - uniform : 같은 인덱스에 해당하는 값을 서로 교차한다.
 - arithmetic : 랜덤 확률로 가중 평균을 구한다.
3. crossover가 적용된 c 개의 토큰 임베딩들 중 mutation 값 m 의 비율만큼 값을 변형시킨다. 매번 5가지 방법 one random, multi random, swap, scramble, inverse 중 하나를 랜덤하게 선택하여 mutation을 진행하며, 각 방법에 대한 설명은 다음과 같다.
- one random : 한 개의 인덱스에 해당하는 값을 랜덤 값으로 바꾼다.
 - multi random : 여러 개의 인덱스에 해당하는 값을 랜덤 값으로 바꾼다.
 - swap : 두 개의 인덱스 값을 서로 바꾼다.
 - scramble : 범위 안에 해당하는 값을 서로 랜덤하게 바꾼다.
 - inverse : 범위 안에 해당하는 값을 반전시킨다.
4. crossover와 mutation이 적용된 c 개의 토큰 임베딩은 다음 iteration에서 토큰 임베딩 후보가 되고, 다음 iteration을 진행한다.
5. 토큰 임베딩 후보 중 각 토큰 임베딩 $S^*:v^*$ 에 대

- 해, T2I 모델의 tokenizer와 text encoder에 저장하고 prompt 'a photo of a S^* '를 T2I 모델에 입력하여 n 개의 이미지를 생성한다. 각 생성 이미지에 대해 objective function에 따라 쿼리 이미지의 임베딩과 생성 이미지의 임베딩 사이의 코사인 유사도를 구하고, 각 토큰 임베딩에서 생성한 모든 이미지에 대한 평균을 계산한다.
6. 토큰 임베딩 후보 중 objective function이 가장 높은 토큰 임베딩을 선택하여, iteration 값 i 만큼 과정 2~5를 반복한다.

본 논문의 Textual Inversion을 활용한 Adversarial Prompt 생성 기반 T2I 모델에 대한 멤버십 추론 공격을 위한 준비(이미지 생성) 과정을 Fig. 1.에 제시하였으며, 과정의 세부사항은 4.1 실험 환경 설정에 설명되어 있다. 본 논문은 생성 이미지를 토대로 threshold(임계값)를 설정하고 멤버십 추론 공격을 진행하였으며, 4.2와 4.3에서 각각 threshold 설정 방법과 멤버십 추론 공격 성능을 확인할 수 있다.

IV. 실험 및 결과

4.1 실험 환경 설정

본 논문은 target T2I 모델로 Stable Diffusion 모델(stable-diffusion-v1-5)을 사용했

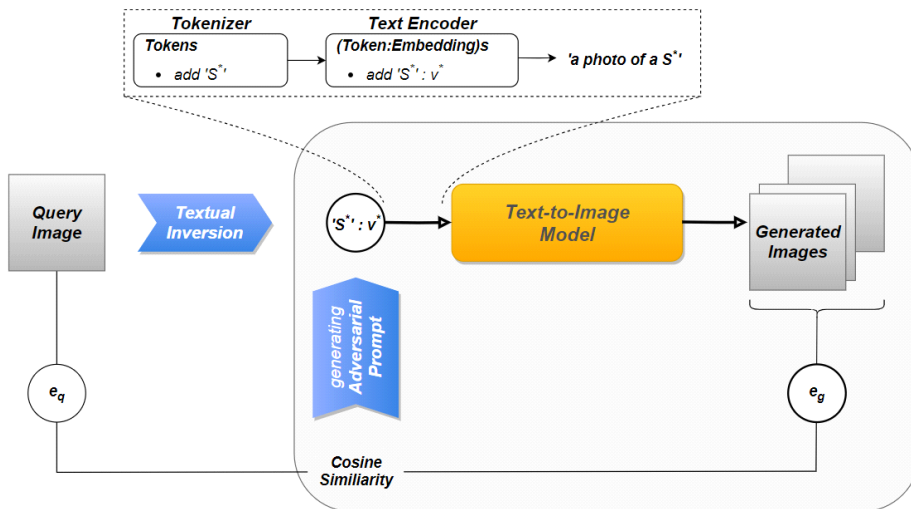


Fig. 1. Process of Proposed Method

Table 2. Setting of Experiments

setting	Textual Inversion	method of generating Adversarial Prompt				generation per 1 query image
	iteration	iteration (i)	generation (n)	crossover (c)	mutation (m)	
[A]	3,000	10	4	2	0.3	40
[B]	3,000	50	10	4	0.3	500
[C]	5,000	50	10	4	0.3	500

으며, Stable Diffusion 훈련에 사용된 LAION dataset[17]을 member dataset으로 사용하였고 훈련에 사용되지 않은 CC12M dataset[18]을 non member dataset으로 사용하였다. 그리고 threshold 설정을 위한 train dataset은 member와 non member 각각 10장의 이미지를 사용하였고, 설정한 threshold로 멤버십 추론 공격을 시행할 test dataset은 member와 non member 각각 20장의 이미지를 사용하였다.

Textual Inversion과 Adversarial Prompt 생성 기반 최적화에 대한 실험 환경 설정을 Table 2.에 나타내었다. Textual Inversion은 쿼리 이미지에 대해 iteration번의 학습을 통한 토큰 임베딩을 사용했다. Adversarial Prompt 생성 기반 최적화로 이미지를 생성할 때는 아래 과정을 iteration(i)번 반복한다. 토큰 임베딩 후보 중 각 토큰 임베딩에 대해 generation(n)개의 이미지를 생성하고, 생성 이미지에 대해 objective function을 계산하여 가장 큰 값을 가지는 토큰 임베딩을 선택한다. 선택한 토큰 임베딩을 crossover(c)개 복제하여 c번의 crossover를 반복하고, 비율 mutation(m)만큼 mutation을 진행하여 토큰 임베딩 후보를 만든다.

4.2 threshold 설정

Textual Inversion과 Adversarial Prompt 생성 기반 최적화에 대한 실험 환경 설정 (A), (B), (C)에 따라 이미지를 생성하였다. 본 논문은 생성 이미지와 쿼리 이미지가 유사할수록 member일 확률이 높다는 직관을 이용하여 멤버십을 추론하므로 생성 이미지에 대한 성능을 계산하였다. 따라서 쿼리 이미지의 임베딩과 생성 이미지의 임베딩 사이의 코사인 유사도(img_img)를 계산하고, pixel 수준의 유사도를 확인하고자 MSE[19], PSNR[20],

SSIM[21]을 계산하였다. 추가로, Carlini 등[16]이 생성 이미지 간 유사도를 통해 memorization을 찾은 것을 이용하여, 생성 이미지 간 성능을 계산하였다. 성능 지표는 쿼리 이미지와 생성 이미지 간 성능과 동일하게, 임베딩 간 코사인 유사도(img_img), MSE, PSNR, SSIM을 계산하였다.

본 논문은 average(평균)와 over(over the average)를 이용하여 threshold를 설정한다. average는 성능의 평균값이고, over는 average 이상의 성능을 갖는 결과를 카운트한 값이다. average와 over 계산은 다음 3가지 범위에 따라 계산된다.

- total : 모든 iteration에 대한 생성 이미지의 average와 over
- iter : 각 iteration에 대한 생성 이미지의 average와 over
- range : 특정 범위(1~5, 5~10, ..., ~iteration)에 해당하는 생성 이미지의 average와 over

본 논문은 각 성능지표(img_img, MSE, PSNR, SSIM)마다 average와 over 중 하나를 threshold 지표로 사용하였다. threshold를 선택하기 위해, 모든 범위(total, iter, range)에 걸쳐 계산한 성능 중에서 member가 non member보다 더 좋은 성능을 갖는 결과를 저장한다. 저장한 결과 중에서 각 성능지표(img_img, MSE, PSNR, SSIM)마다 카운트가 많은 average 또는 over가 threshold 지표가 되고, 선택된 threshold 지표에 저장된 결과의 평균값을 threshold 값으로 사용한다. 설정한 threshold 예시를 Table 3.에 나타내었다. 쿼리 이미지에 대해 Textual Inversion과 Adversarial Prompt 생성 방법으로 생성한 이미지의 성능(img_img, MSE, PSNR, SSIM) 중에서 설정한 threshold 중 하나라도 만족을 하면, 해

당 쿼리 이미지를 member라고 예측한다.

4.3 실험 결과

Textual Inversion과 Adversarial Prompt 생성 기반 최적화로 이미지를 생성하고, 생성 이미지의 성능을 계산하여 설정한 threshold로 멤버십 추론 공격을 진행하였다. 제안 방법에 대한 멤버십 추론 공격의 성능을 Table 4.에 나타내었다. 성능 지표는 3가지를 사용했으며, 다음과 같다.

- TPR (True Positive Rate) : member를 member로 맞게 예측한 비율
- TNR (True Negative Rate) : non member를 non member로 맞게 예측한 비율
- Accuracy : 전체 중에서 member와 non member를 맞게 예측한 비율

공격 성능은 적은 확률(0.5)보다 높은 최소 0.55에서 최대 1.00의 공격 Accuracy를 달성하였으며, iter와 range에 대한 성능보다 total에 대한 성능이 훨씬 좋다. 이는 threshold를 모든 범위(total, iter, range)에 대한 결과(average, over)를 기반으로 설정하므로, threshold를 설정한 결과의 분포와 모든 iteration을 다루는 total의 결과(average, over) 분포와 비슷해지기 때문인 것 같다.

Table 3. Example of Threshold

img_img		MSE		PSNR		SSIM	
avg ↑	over ↑	avg ↓	over ↑	avg ↑	over ↑	avg ↑	over ↑
-	45.98	0.12	-	10.01	-	0.25	-

Table 4. Performances of Membership Inference Attack

setting	measure	query img~generated img			generated img~generated img		
		total	iter	range	total	iter	range
[A]	TPR	1.00	1.00	1.00	1.00	0.40	1.00
	TNR	0.40	1.00	0.10	0.90	1.00	0.10
	Accuracy	0.70	1.00	0.55	0.95	0.70	0.55
[B]	TPR	1.00	0.55	1.00	1.00	0.45	1.00
	TNR	0.95	1.00	0.50	0.90	1.00	0.45
	Accuracy	0.97	0.78	0.75	0.95	0.72	0.72
[C]	TPR	1.00	0.60	1.00	1.00	0.40	1.00
	TNR	0.80	1.00	0.50	0.85	1.00	0.50
	Accuracy	0.90	0.80	0.75	0.93	0.70	0.75

V. 결론

본 논문은 T2I 모델에 대한 멤버십 추론 공격의 기존 연구[1]보다 공격 성능을 높이기 위한 시도로 Textual Inversion과 Adversarial Prompt 생성 방법을 활용하였다. T2I 모델에 대한 멤버십 추론 공격 관점에서 caption으로 이미지를 생성할 때 쿼리 이미지를 잘 반영하지 못하는 문제점을 발견하고, T2I 모델에 대한 personalization인 Textual Inversion으로 쿼리 이미지를 표현함으로써 이를 보완하였다. 또한, 여러 장의 이미지를 효과적으로 생성하기 위해 Adversarial Prompt 생성 방법을 활용하여, 쿼리 이미지와 유사한 이미지가 생성될 수 있도록 objective function을 정의하고 T2I 모델의 입력력을 최적화하였다. 또한, 최초로 Stable Diffusion에 대해 멤버십 추론 공격을 진행하였으며, 최대 1.00의 Accuracy를 달성하였다.

향후에는 본 논문에서 제안한 멤버십 추론 공격에 대한 이론적 타당성을 입증하기 위해, 실험을 위한 dataset을 확장하고 더 다양한 분석을 진행할 계획이다.

References

- [1] Y. Wu, N. Yu, Z. Li, M. Backes, and Y. Zhang, "Membership inference attacks against text-to-image generation models," arXiv preprint arXiv:2210.00968, Oct. 2022.
- [2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," In Proceedings of the IEEE/CVF Conference on Computer vision and Pattern Recognition, pp. 10684-10695, Jun. 2022.
- [3] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," arXiv preprint arXiv:2204.06125, 1(2), Apr. 2022.
- [4] J. Hayes, L. Melis, G. Danezis, and E.D. Cristofaro, "LOGAN: Membership inference attacks against generative models," arXiv preprint arXiv:1705.07663, May. 2017.
- [5] K.S. Liu, C. Xiao, B. Li, and J. Gao, "Performing co-membership attacks against deep generative models," In Proceedings of the 19th IEEE International Conference on Data Mining, pp. 459-467, Nov. 2019.
- [6] B. Hilprecht, M. Härterich, and D. Bernau, "Monte carlo and reconstruction membership inference attacks against generative models," Proceedings on Privacy Enhancing Technologies, pp. 232-249, Jul. 2019.
- [7] D. Chen, N. Yu, Y. Zhang, and M. Fritz, "GAN-Leaks: A taxonomy of membership inference attacks against generative models," In Proceedings of the 2020 ACM SIGSAC conference on computer and communications security, pp. 343-362, Nov. 2020.
- [8] N. Maus, P. Chao, E. Wong, and J. Gardner, "Adversarial prompting for black box foundation models," arXiv preprint arXiv:2302.04237, May. 2023.
- [9] H. Zhuang, Y. Zhang, and S. Liu, "A pilot study of query-free adversarial attack against stable diffusion," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2384-2391, Jun. 2023.
- [10] Q. Liu, A. Kortylewski, Y. Bai, S. Bai, and A. Yuille, "Intriguing properties of text-guided diffusion models," arXiv preprint arXiv:2306.00974, Nov. 2023.
- [11] H. Liu, Y. Wu, S. Zhai, B. Yaun, and N. Zhang, "RIATIG: Reliable and imperceptible adversarial text-to-image generation with natural prompts," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20585-20594, Jun. 2023.
- [12] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A.H. Bermano, G. Chechik, and D. Cohen-Or, "An image is worth one word: personalizing text-to-image generation using textual inversion," arXiv preprint arXiv:2208.01618, Aug. 2022.
- [13] Z. Fei, M. Fan, and J. Huang, "Gradient-free textual inversion," arXiv preprint arXiv:2304.05818, Apr. 2023.
- [14] R. Gal, M. Arar, Y. Atzmon, A.H. Bermano, G. Chechik, and D. Cohen-Or, "Encoder-based domain tuning for fast personalization of text-to-image models," ACM Transactions on Graphics (TOG), vol. 42, no. 4, pp. 1-13, Aug. 2023.
- [15] L. Melas-Kyriazi, I. Laina, C. Rupprecht, and A. Vedaldi, "Realfusion: 360° reconstruction of any object from a single image," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.8446-8455, Jun. 2023.
- [16] N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Sehwal, F. Tramèr, B.

- Balle, D. Ippolito, and E. Wallace, "Extracting training data from diffusion models," In 32nd USENIX Security Symposium (USENIX Security 23), pp. 5253-5270, Aug. 2023.
- [17] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, J. Jitsev, et al, "Laion-5b: An open large-scale dataset for training next generation image-text models," Advances in Neural Information Processing Systems, Nov. 2022.
- [18] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 25278-25294, Jun. 2021.
- [19] A. Botchkarev, "Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology," arXiv preprint arXiv:1809.03006, 2018.
- [20] A. Horé and D. Ziou, "Image quality metrics: PSNR vs. SSIM", 20th international conference on pattern recognition. IEEE, pp. 2366-2369, Aug. 2010.
- [21] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE transactions on image processing, vol. 13, no. 4, pp.600-612, Apr. 2004.

 < 저자 소개 >



오 윤 주 (Yoonju Oh) 학생회원
 2021년 2월: 공주대학교 응용수학과 학사
 2022년 2월~현재: 숭실대학교 소프트웨어학과 석사과정
 <관심분야> 개인정보보호, 연합학습, 생성 모델, AI 보안



박 소 희 (Sohee Park) 학생회원
 2018년 2월: 공주대학교 응용수학과 학사
 2020년 2월: 공주대학교 융합과학과 석사
 2020년 6월~2021년 9월: 한국교육학술정보원 전문원
 2022년 2월~현재: 숭실대학교 소프트웨어학과 박사과정
 <관심분야> 금융보안, 인공지능 보안



최 대 선 (Daeseon Choi) 종신회원
 1995년 2월: 동국대학교 컴퓨터공학과 학사
 1997년 2월: 포항공과대학교 컴퓨터공학과 석사
 2009년 1월: 한국과학기술원 전산학과 박사
 1997년 1월~1999년 6월: 현대정보기술 선임
 1999년 7월~2015년 8월: 한국전자통신연구원 인증기술연구실 실장/책임연구원
 2015년 9월~2020년 8월: 공주대학교 의료정보학과 부교수
 2020년 9월~현재: 숭실대학교 소프트웨어학부 교수
 2016년~현재: 정보보호학회 이사
 <관심분야> 인증, 개인정보보호, AI 보안

