2025 아인슈타인 & 피카소 창의융합 오디세이

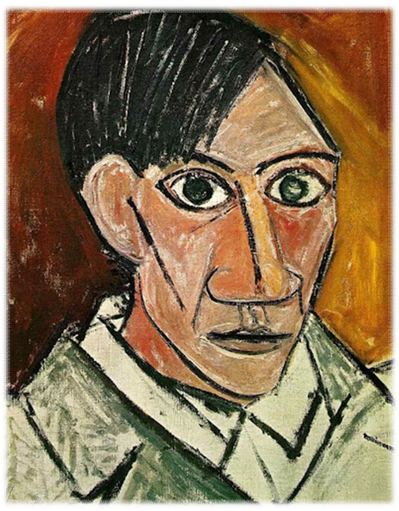
피카소 프로토콜: 비대칭 신경망 모델을 활용한 공개키 스테가노그래피

GitHub

이은세(1308)

서론

파블로 피카소는 입체주의(Cubism)의 창시자이자 20세기 미술의 상징이다. 그는 전통적 구도와 시점을 파괴하고, 사물을 여러 각도에서 동시에 표현하는 기법으로 예술의 새로운 지평을 열었다. 본 프로젝트 “비대칭 스테가노그래피”는 그의 ‘형태 해체와 재구성’ 철학에서 영감을 받았다. 본 연구는 텍스트를 추상 이미지로 변환하고 다시 복원하는 과정을 통해 새로운 데이터 보호 체계를 제시한다.

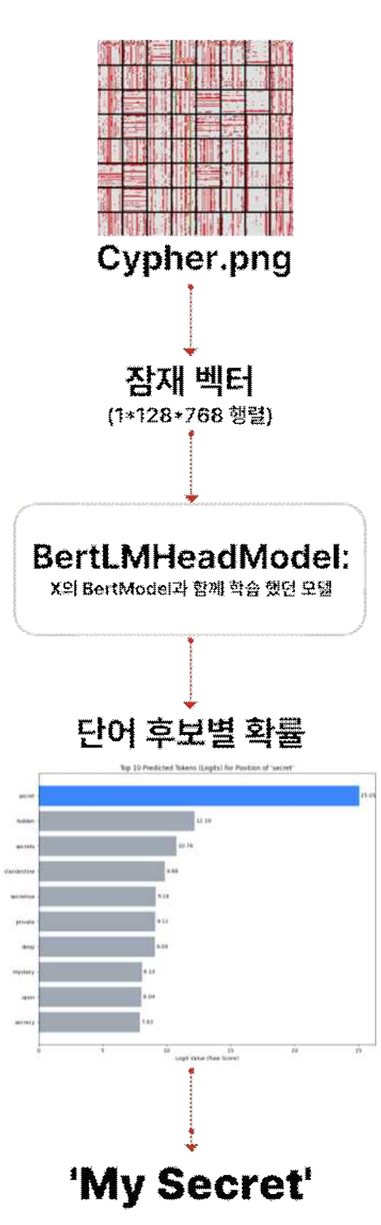


피카소 자화상

1. 핵심 기능 메커니즘

Artist Y(enc):

Artist X(dec):



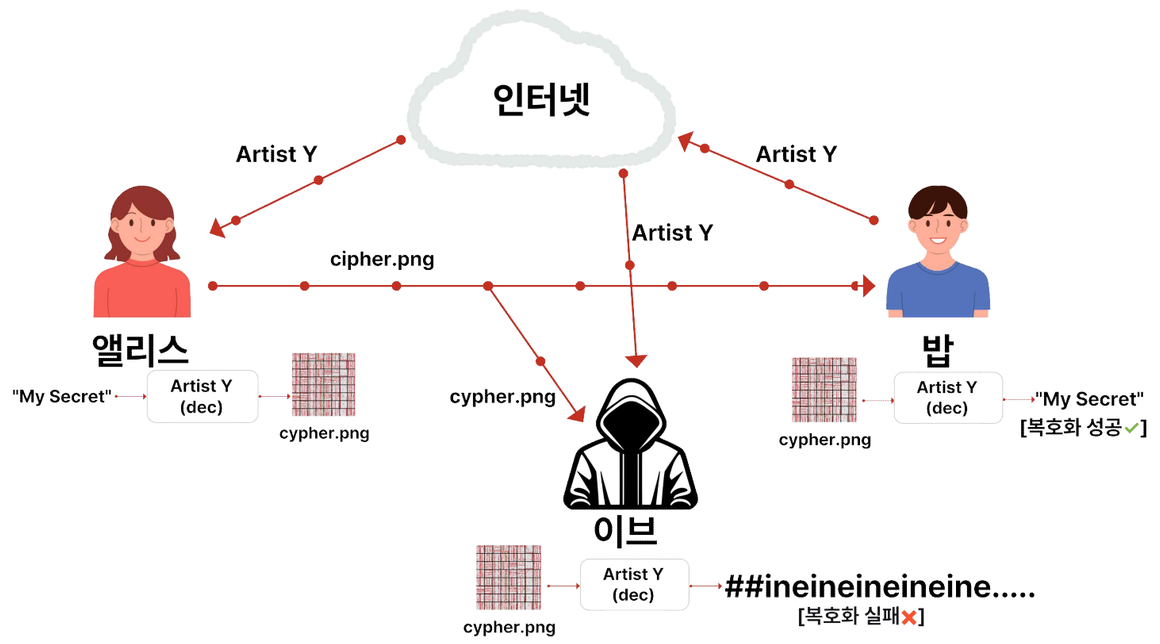
잠재 벡터: 모델만의 사고 과정

BertLM(Bert Language Model): 자연어 처리용 사전학습 언어모델

BertLMHeadModel: BertLM과 함께 학습해 BertLM의 잠재 벡터를 읽을 수 있는 모델

2. 전송 과정

수신자 밥은 자신의 Artist X와 쌍을 이루는 Artist Y를 인터넷에 공개하고, 송신자 앨리스는 이 Artist Y로 평문을 암호 이미지로 변환해 전송하며, 밥은 비밀키인 Artist X로 이를 해독해 평문을 얻는다. 해커 이브는 공개된 Artist Y와 암호 이미지를 입수하더라도 비공개인 Artist X가 없으므로 복호화할 수 없다.



3. 보안성 확인

위와 같이 공개 신경망인 Artist Y의 디코더로는 복호화할 수 없다. 따라서 Artist Y의 인코더를 활용해 Artist X의 디코더를 역추론하는 방법이 가장 유력해 보인다. 이하에서는 해당 시나리오가 불가능함을 수식을 통해 증명한다.

1. 탐색 공간

디코더의 파라미터 개수 $\approx 1.1 \times 10^8$

각 파라미터(32비트 부동 소수점)의 가능한 값의 개수: 2^{32}

따라서 나올 수 있는 경우의 수: $2^{32 \times (1.1 \times 10^8)} \approx 10^{1.06 \times 10^9}$

2. 동일 데이터 셋으로 모델 생성

Artist X(dec)의 BertModel은 학습 시마다 wikitext-2-raw-v1(위키백과 텍스트 데이터 셋)의 일부를 무작위로 제외한다. 따라서 어떤 데이터가 제외되었는지 확인할 수 없으므로, 동일한 데이터 셋을 재현하는 것은 불가능하다.

4. 결론 및 향후 연구

본 프로젝트는 피카소의 예술 철학에서 영감을 받아 비대칭 신경망 기반 스테가노그래피 시스템을 설계·구현하였다. 향후 과제로는 증명할 수 있는 보안 확보를 제안한다. 이를 위해 데이터 은닉 여부를 판별하는 신경망과 이를 속이는 인코더를 적대적 학습으로 함께 훈련해 통계적으로 탐지 불가능함을 입증한다면, 본 시스템의 보안성은 계산적 어려움 기반에서 수학적으로 증명할 수 있는 수준으로 격상될 것이다.