

동형암호를 이용한 프라이버시 보호 얼굴 검증 시스템

심준석, 조현진, 홍윤영, 김호원*

*부산대학교

sugo312@pusan.ac.kr, wh77r77@pusan.ac.kr, hyy0238@pusan.ac.kr, *howonkim@pusan.ac.kr

Privacy Protection Face Verification System using Homomorphic Encryption

Junseok Shim, Hyunjin Jo, Yoonyoung Hong, HowonKim*

*Pusan Univ.

요약

최근 딥러닝 기술의 발전에 따라 컴퓨터 비전 분야에서 얼굴 검증 기술에 대한 다양한 연구가 제안되고 있다. 그러나 딥러닝 모델 훈련에 사용되는 데이터를 추출하는 공격인 모델 전도 공격에 의해 얼굴 검증 시스템에 등록된 이미지가 유출될 가능성이 있어 프라이버시 침해 문제가 제기되고 있다. 본 논문에서는 프라이버시 침해 문제를 해결하기 위해 Paillier 동형암호와 삼 네트워크를 사용한 얼굴 검증 시스템을 제안한다. 이미지에 Paillier 동형암호화를 수행하고 삼 네트워크를 통해 암호화 이미지 사이의 얼굴 검증을 수행한다. ROC 커브를 통해 원본 이미지의 얼굴 검증 결과와 암호화 이미지의 얼굴 검증 결과를 비교하였다. 실험 결과, 원본 이미지와 암호화 이미지의 얼굴 검증 결과로 비슷한 AUC를 보였으며, 프라이버시 보호와 동시에 얼굴 검증이 가능함을 보였다.

I. 서론

얼굴 검증은 입력으로 들어오는 두 개의 얼굴 영상이 동일 인물인지 여부를 판단하는 1:1 검증 문제이다. 최근 딥러닝 기술의 발전으로 딥러닝 기술을 적용시킨 얼굴 검증(Face verification) 기술의 다양한 연구가 제시되고 있으며, FaceID 등 모바일 앱 로그인 및 결제 시스템에서 사용되고 있다. 그러나 딥러닝을 활용한 보안기술 연구가 활발히 이루어지는 가운데, 딥러닝을 악용한 연구사례들도 많이 나타나고 있다. 모델 전도 공격(Model inversion attack)은 딥러닝 모델의 학습에 사용된 데이터를 추출하는 공격기법이다. 주어진 입력의 출력 결과와 신뢰도를 분석하여 역으로 데이터를 추출해 낼 수 있다. 얼굴 검증 모델의 경우, 공격자가 모델의 학습에 사용된 이미지를 추출하여 프라이버시 침해 문제를 초래할 수 있다. 따라서 이미지가 유출되더라도 피해를 최소화하기 위해 데이터 암호화가 필요하다. Munirah 등[1]은 인공위성으로 촬영한 이미지를 보호하기 위해 동형암호를 적용한 모델 학습 기법을 제안했다. Ma 등[2]은 얼굴의 특징을 통해 등록 및 검증 시스템인 DeepID를 제안하였다. DeepID 시스템은 Convolutional Neural Network(CNN) 기반으로 추출된 얼굴 특징을 암호화 모듈을 통해 암호화한다. 시스템에 저장된 특징들과 암호화된 특징 간의 해밍 거리(Hamming distance) 값이 임계값 이하일 경우 동일 인물이라고 판별하는 시스템이다.

본 논문에서 제안하는 시스템은 암호화를 통해 얼굴 이미지 데이터의 프라이버시를 보호하면서 얼굴 검증을 한다. 또한 [2]와 달리 기존의 CNN 기반의 얼굴 검증 시스템이 아닌 삼 네트워크(Siamese network)를 사용해 소량의 이미지만으로 얼굴 검증 결과에 높은 성능을 보인다.

본 시스템에서는 프라이버시 보호 및 얼굴 검증을 위해 Paillier 암호 체계와 삼 네트워크를 사용했다. Paillier 암호 체계는 비밀키에 대한 정보가 없더라도 암호문만 이용하여 평문 간의 상수 곱과 덧셈 결과를 구할 수 있는 동형 연산이 가능한 특징을 가진다. 이미지의 동형암호화는 그림 2와 같이 먼저 이미지를 1차원 배열로 변환한 후 공개키를 통해 암호화한다. 암호화된 1차원 형태의 배열은 모델에 입력하기 위해 이미지의 원래 모양으로 변형하여 암호화를 완료한다.

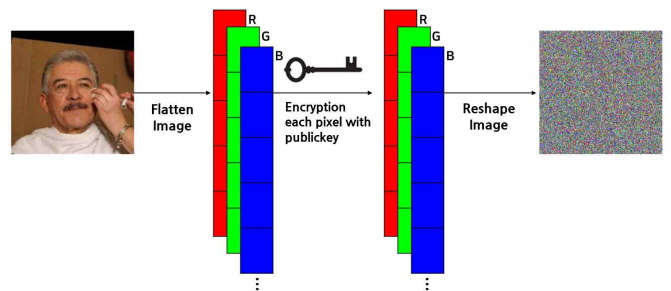


그림 1 이미지 암호화 과정

삼 네트워크는 그림 2와 같이 동일한 파라미터를 공유하는 트윈 네트워크(Twin network)를 포함하는 딥러닝 모델이다. 트윈 네트워크는 검증하고자 하는 이미지와 검증용 참조 이미지로 구성된 이미지 쌍을 입력받아 특징 벡터를 추출한다. Distance layer에서 추출된 특징 벡터들의 L_1 거리를 계산한 후, Prediction layer에서 L_1 거리를 이용해 유사도를 계산한다.

2.2 모델 구조

본 논문에서 제안하는 얼굴 검증 시스템의 삼 네트워크 모델은 $128 \times 128 \times 3$ 크기의 이미지 쌍을 입력받는다. 이미지 쌍은 트윈 네

II. 본론

2.1 배경 지식

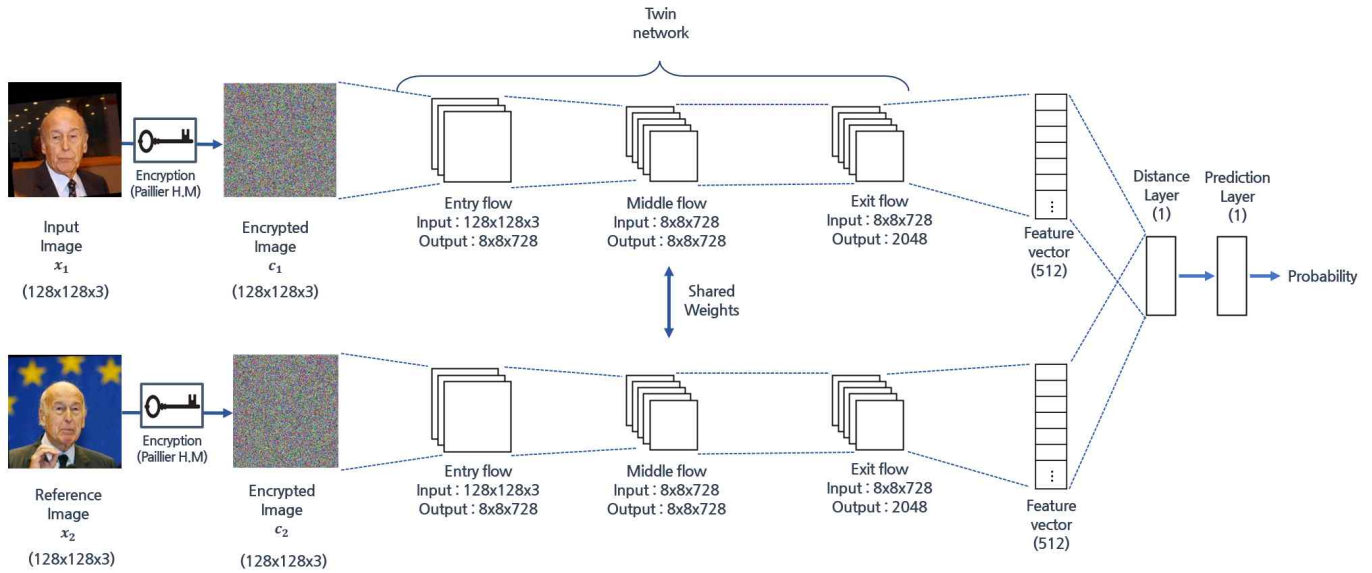


그림 2 시스템 구조

트윈네트워크에 입력되기 전 동형암호화를 거친다. 본 논문에서는 연산의 효율을 위해 ImageNet[3] 데이터 세트에서 사전 학습된 XceptionNet[4]을 트윈 네트워크로 사용한다. 삼 네트워크의 Prediction Layer에서 출력되는 입력 이미지 쌍의 유사도는 입력 이미지의 인물들을 동일 인물로 판단하는 기준이 된다.

2.3 실험 및 실험 결과

본 연구에서는 얼굴 검증 모델의 학습을 위해 Labeled Faces in the Wild(LFW)[5] 데이터 세트를 사용한다. LFW 데이터 세트는 5,749명의 사람에 대해 13,233개의 이미지로 구성된다. 학습 및 검증을 위해 LFW 데이터 세트에 존재하는 1,680명에 대해 동일 인물로 구성된 이미지 쌍과 서로 다른 인물로 구성된 이미지 쌍으로 데이터를 구성했다. 동일 인물로 구성된 이미지 쌍은 1, 서로 다른 인물로 구성된 이미지 쌍에는 0의 레이블을 부여했다.

시스템의 성능 평가를 위해 원본 이미지와 암호화 이미지의 얼굴 검증 결과의 Receiver Operating Characteristic(ROC) 커브를 사용한다. 그림 3은 원본 이미지와 암호화 이미지의 ROC 커브 비교이다. 원본 이미지와 암호화 이미지의 AUC(Area Under the ROC curve)는 각각 0.99, 0.97이다. Park[8]에 따르면 AUC에 따른 정확도는 AUC가 0.5~0.7 일 때 낮은 정확도(less), 0.7~0.9 일 때 중등도(moderate), 0.9~1.0 일 때 높은 정확도(high)를 가진다. 따라서 암호화 이미지에 대한 얼굴 검증 결과와 원본 이미지의 얼굴 검증 결과가 큰 차이가 없이 높은 정확도의 결과를 가지는 것을 확인할 수 있다.

III. 결론

본 논문에서는 얼굴 이미지에 대한 프라이버시를 보호하는 얼굴 검증 시스템을 제안하였다. 제안된 기술은 Paillier 동형암호를 사용해 암호화된 이미지를 학습 및 검증에 사용해 모델 진도 공격으로 인해 유출되는 이미지에 대하여 프라이버시를 보호한다. 또한 삼 네트워크를 통해 적은 암호화 이미지 데이터에 대한 학습이 가능함을 보였다. 향후, 본 솔루션을 통해 얼굴 검증 이외 프라이버시 보호에 민감한 객체 탐지 분야 및 컴퓨터 비전 분야에서도 사용될 것으로 전망한다.

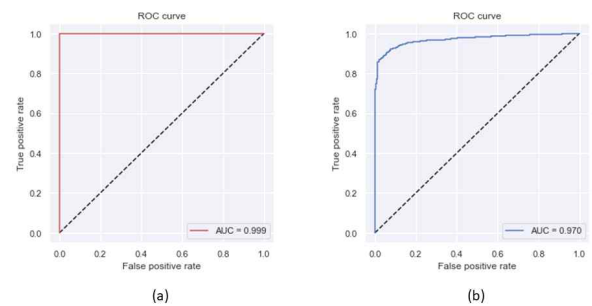


그림 3. ROC Curve 비교 (a) 원본 이미지의 ROC Curve, (b) 암호화 이미지의 ROC Curve

ACKNOWLEDGMENT

이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (2019-0-01343, 융합보안핵심인재양성사업)

참고 문헌

- [1] Alkhelaiwi, Munirah, et al. "An Efficient Approach Based on Privacy-Preserving Deep Learning for Satellite Image Classification." *Remote Sensing* 13.11 (2021): 2221.
- [2] Ma, Yukun, et al. "A secure face-verification scheme based on homomorphic encryption and deep neural networks." *IEEE Access* 5 (2017): 16532-16538.
- [3] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems* 25 (2012): 1097-1105.
- [4] hollet, François. "Xception: Deep learning with depthwise separable convolutions." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
- [5] Park, S.-I., & O, T.-H. (2016). Application of Receiver Operating Characteristic (ROC) Curve for Evaluation of Diagnostic Test Performance. *Journal of Veterinary Clinics*, 33(2), 97. <https://doi.org/10.17555/jvc.2016.04.33.2.97>