인공신경망을 이용한 조작된 신분증 탐지기법 연구

문학준(학부생), 박은주, 김정호, 윤관식*, 서연아*, 우사이먼성일

성균관대학교 소프트웨어학과

Manipulated ID Card Classification using Deep Neural Networks

Hakjun Moon, Eunju Park, Jeongho Kim, Kwansik Yoon, Yeonah Seo, Simon S. Woo

Department of Computer Science and Engineering, Sungkyunkwan University

Samsung SDS*

요 약

전자상거래와 디지털 뱅킹 위주로 모바일 신원 인증 시스템이 많이 사용되고 있다. 비대면으로 서비스를 이용하기 위해 사용자의 신원을 인증하는 과정에서 주민등록증이나 운전면허증과 같은 신분증을 촬영하게 된다. 그러나 사용자의 카메라로는 실물 신분증을 촬영하고 있음을 확인할 수 없으므로 촬영된 신분증의 진위여부를 파악할 필요가 있다. 이 논문에서는 사용자가 원격으로 제공하는 신분증 이미지가 실제인지, 비디지털 영역(고화질로 인쇄된 이미지 또는 촬영 후 모니터에 출력된 이미지)에서 조작되었는지를 판별하기 위해 딥러닝 기법을 사용하였다. 모델의 입력으로 RGB 이미지 외에도 이산 푸리에 변환과 특징 추출 기법을 사용하여 실험하였다. 학습한 모델을 이용하여 신분증 이미지의 진위여부를 판별하였을 때 최대 96.6%의 분류 정확도를 달성하였다.

I. 서론

디지털 기술의 급속한 발전은 사회의 많은 부분을 변화시켰다. 특히, 전자상거래와 디지털 뱅킹 등의 온라인 서비스는 이러한 전환의 중심에 있으며, 이런 서비스들은 사용자의 생활을 편리하게 하고 경제 활동을 활성화시키는 데 크게 기여하고 있다. 하지만, 이러한 변화는 새로운 문제점들을 수반한다. 그 중 하나가 바로 신원 인증의 문제이다. 비대면 서비스에서는 주로 주민등록증이나 운전면허증 등의 신분증을 통해 사용자의 신원을 인증하게 된다. 하지만 사용자가 직접 카메라로 신분증을 촬영해 제출하는 경우, 그 신분증이 진짜인지, 아니면 인쇄했거나 모니터에 있는 이미지를 사용했는지 빠르고 정확하게 판별하는 것은 어려운 일이다. 또한, 최근 전 세계적으로 코로나 19로 인한 팬데믹 상황이 장기화함에 따라, 비대면 서비스 제공의 필요성이 대두되었다. 이러한 상황에서 사용자의 신원을 안전하게 확인할 수 있는 방법은 더욱 중요하게 되었다. 이에 따라, 신분증의 진위를 정확하게 판별할 수 있는 기술이 필요하다.

문제에 대한 해결 방법으로 우리는 딥러닝, 특히 합성곱 신경망 (Convolutional Neural Network, CNN) [1] 기반의 신분증 진위 판별 방법을 제시하려고 한다. CNN은 이미지 내의 복잡한 패턴을 학습하여 고정된 규칙을 기반으로 한 전통적인 이미지 분석 방법보다 더욱 높은 성능을 보이고 있다. 이를 신분증 진위 판별에 적용하면, 실물 여부와 디지털/비디지털 영역에서의 조작 여부를 알아내는 것이 가능하다.

RGB 이미지 외에도 이산 푸리에 변환(Discrete Fourier Transform, DFT)과 오차 수준 분석(Error Level Analysis, ELA)를 사용하여 신분증 이미지의 특징을 추출하였다. DFT는 신호 처리에서 주로 사용되는 기법으로, 복잡한 신호를 여러 개의 단순한 주파수 성분으로 분해하여 분석하는 방법이며, ELA는 디지털 이미지의 압축 오차를 분석하여 이미지가 수정된 정도를 판별하는 기법이다. 이 기법을 신분증 이미지 분석에 적용했을 때 신분증 이미지가 디지털적으로 조작되었는지를 판별할 수 있다고 생각하였다. CNN을 이용하여 촬영된 신분증의 실물, 종이 출력, 모니터 출력 여부를 분류하는 실험을 RGB raw 이미지를 진행하였으며, 높은 정확도로 판별할 수 있음을 보인다. 또한 DFT, ELA의 특징 추출 방법이 신분증 사진의 진위 여부 판별에 사용될 수 있는지 실험하였다.

본 논문에서는 결론적으로 딥러닝을 포함한 인공지능이 신원 인증과 같은 중요한 문제 해결에 큰 역할을 할 수 있음을 제시한다. 이를 활용하면 서비스 제공자와 사용자 모두에게 강화된 보안성과, 향상된 편의성을 제공할 수 있을 것이다. 이 논문을 통해 보다 신뢰할 수 있는 디지털 세상을 만드는 데 일조하고자 한다.

II. 이론적 배경 및 관련 연구

2.1 모델 설명 및 이론적 배경

본 연구에서는 실험 모델로 EfficientNet(b1, b3), MobileNetV2를 사용하였으며, RGB raw 이미지 외에 이산 푸리에 변환(DFT)과 오차 수준 분석(ELA) 기법을 사용하여 이미지로부터 특징을 추출하였다.

1) EfficientNet

EfficientNet [2]은 2019년 Google에서 개발한 CNN 모델이다. 기존의 CNN Architecture에서 발전시킨 개념인 Compound Scaling을 사용한다. 모델의 capacity를 증가시키는 Depth scaling, Width scaling, Resolution scaling의 세 가지 요소를 각각 단독으로 조정하는 것보다, 조화롭게 조정하면 더 나은 성능을 얻을 수 있다는 것이 EfficientNet의 주요 아이디어이다. 네트워크의 깊이, 너비, 이미지 해상도에 따라 b0~b7의 버전이 있으며, 본 논문에서는 작은 모델 간 성능을 비교하기 위하여 b1과 b3을 사용하였다.

2) MobileNetV2

MobileNetV2 [3]는 Google이 개발한 경량 CNN으로, 특히 휴대폰, 태블릿, IoT 장치 등과 같은 edge 컴퓨팅 환경과 임베디드 시스템 등 계산 능력이 제한된 환경에서의 높은 성능을 목표로 한다. MobileNetV2은 Inverted Residuals와 Linear Bottlenecks라는 두 가지 기법을 사용한다. Inverted Residuals는 기존 ResNet의 구조를 '뒤집어' 적용한 것으로, 좁은 차원에서 넓은 차원으로 정보를 전달한다. 이로 인해 계산 복잡도를 줄이면서도 네트워크가 필요한 정보를 유지한다. 반면 Linear Bottlenecks은 각 블록의 출력에서 ReLU와 같은 비선형 활성화 함수를 제거하여 정보의 손실을 최소화한다. 이러한 설계는 모델이 충분한 표현력을 유지하면서도 전체 모델 크기와 연산 복잡도를 크게 줄이는 데 도움을 준다.

2.2 이미지 특징 추출 기법

1) 이산 푸리에 변환(Discrete Fourier Transform, DFT)

이산 푸리에 변환은 복잡한 신호를 주파수 영역으로 변환하는 방법으로, 이를 통해 신호의 주파수 구성 요소를 분석할 수 있다. 이 기법은 시간에 따라 변화하는 신호를 주파수 성분으로 분해하여, 고주파와 저주파 성분을 구분한다. 이미지에서의 이산 푸리에 변환은 이미지 내부의 다양한 패턴과 질감을 분석하는 데 유용하다. 예를 들어, 고주파 성분은 이미지의 세부 정보나 경계선을 나타내며, 저주파 성분은 이미지의 전반적인 빛과 그림자, 색상 등을 나타낸다. 이러한 특성은 이미지에서 복잡한 특징을 추출하는 데에 유용하며, 본 연구에서는 신분증 이미지에 이산 푸리에 변환을 사용하였다.

2) 오차 수준 분석(Error Level Analysis, ELA)

오차 수준 분석은 디지털 이미지가 압축 과정에서 발생하는 특정한 오차를 분석하는 기법이다. JPEG 이미지를 저장할 때에 압축하는 과정을 포함하는데, 이 과정에서 이미지의 상세한 정보가 손실되며, 이 때 생성되는 오차를 분석하는 것이 ELA이다.

이 기법은 이미지의 각 부분이 원래의 상태와 얼마나 다른지를 측정하여, 이미지가 얼마나 일관성 있게 압축되었는지를 판단한다. 이미지의 특정 부분이 다른 부분에 비해 더 많은 오차를 보이면, 그 부분은 원본에서 수정되었을 가능성이 높다. 이러한 특성은 이미지의 위조 여부를 판별하기 위한 특징 중 하나로 사용될 수 있다.

2.3 관련 연구

이 장에서는 신분증, 지문, 얼굴 인식을 중심으로 한 원격 인증 시스템에 대한 기존 연구를 소개한다.

Stokkenes et al. [4]은 bloom filter를 사용하여 얼굴에서 추출한 특징을 기반으로 한 온라인 뱅킹 인증 시스템을 제안하였으며, Shi et al. 은 신분증 사진과 selfie 사진을 대조하는 시스템인 DocFace[5]와 후속 연구인 DocFace+ [6]를 제안하였다. 이 연구는 transfer learning 기법을 이용하여 얼굴인식 baseline 모델을 selfie 데이터셋에 맞게 학습하였다.

Pepera et al. 은 사용자의 초기 인증 이후에도 지속적으로 사용자의 신원을 확인하는 active authentication system [7]을 제안하였다.

Benalcazar et al. 은 칠레 신분증 데이터셋의 실물과 합성, 모니터에 출력된 신분증을 CycleGAN, Style-GAN2 등의 모델로 생성하는 연구를 진행하였으며 [8], S. Gonzalez et al. 은 칠레 신분증 데이터셋으로 실물 신분증과 composite(합성)된 데이터, 그리고 실물 신분증, 출력된 신분증, 모니터에 나타난 신분증 데이터를 구분하는 2-stage 모델을 제안하였다. [9] 우리는 이 연구에서 사용한 DFT와 ELA 기법을 한국 신분증 데이터셋에 적용하기로 하였다.

III. 데이터셋

3.1 데이터셋 수집

본 연구에서 사용되는 신분증 데이터셋은 한국의 실제 신분증 및 그 규격을 따르는 가짜 신분증으로 이루어져 있다. 여기에서 실제 신분증이란 한국에서 실제로 본인 확인을 위한 수단으로 사용되는 신분증을 말하며, 자의적 참여의사를 밝힌 본 연구 참여자 및 그 주변인들의 주민등록증 또는 운전면허증을 사용하였다. 또한, 신분증의

다양한 상황을 묘사하고자 가짜 신분증을 제작하였다. 그림 1에서 나타낸 것처럼, 가짜 신분증은 실제 신분증의 규격을 따르며 신분증에 표시된 이미지와 글씨 정보는 모두 실제로 존재하지 않는 사람을 표현하였다. 가짜 신분증의 제작방법 및 본 연구의 개인정보 수집과 이용에 관한 모든 법적 검토는 삼성 SDS에서 대리하였으며, 참여 연구원들은 절차에 맞는 교육을 이수하였다. 실제 신분증 및 가짜 신분증을 사용한 데이터셋 수집 과정에서 이미지 촬영 시 사용한 핸드폰은 아이폰 12 프로, 갤럭시 S22, 갤럭시 노트20, 갤럭시 s21이다. 모니터는 삼성 c27jg54, BenQ GW2780이며, 프린터는 HP의 컬러 레이저젯 프로 MFP M477fdw를 사용하였다. 신분증 이미지는 모두 금융정보회사에서 서비스 사용자의 본인인증 용도로 사용하는 어플리케이션을 사용하여 촬영하였다. 촬영된 이미지는 본인인증 어플리케이션에서의 자체 이미지 프로세싱 외에 다른 포스트 프로세싱을 거치지 않았으며, 촬영환경에 따라 다양한 해상도를 가진다. 또한 다양한 실제상황을 묘사하기 위해 실내와 실외, 신분증의 배경, 촬영각도, 사물과 카메라의 거리 등의 규격을 정하여 촬영하였다.

(1) 위조된 신분증 실물을 촬영한 이미지

(2) 위조된 신분증을 모니터상에서 촬영한 이미지

(3) 위조된 신분증을 종이에 출력한 후 촬영한 이미지

[그림 1] 서로 다른 레이블을 가진 이미지의 예시

3.2 데이터셋 분류

본 연구에서는 신분증 이미지의 진위를 판별하기 위해 다음과 같은 3가지의 경우를 고려한다. 비대면 서비스의 사용자가 실물 신분증을 촬영하는 경우, 종이에 출력된 신분증을 촬영하는 경우, 모니터에 출력된 신분증을 촬영하는 경우이며, 각각 genuine, print, screen의 3가지 레이블로 표현된다. 이 때, 실물 신분증에 대한 이미지에 대한 경우만 real 클래스로 분류되며, 종이 혹은 모니터에 출력된 신분증의 이미지는 fake 클래스로 분류된다.

IV. 실험 결과 및 분석

4.1 실험 방법 및 실험 결과

본 연구에서는 성능 비교를 위하여 세 가지 CNN 모델인 EfficientNet_b1, EfficientNet_b3, MobileNetV2를 이용하여 실물 신분증, 종이에 출력된 신분증, 그리고 촬영한 후 모니터에 표시된 신분증 이미지를 분류하는 실험을 수행하였다.

3종류의 다른 실험으로 구성하였으며, 각 실험에서 모델의 입력으로 사용되는 이미지의 종류를 달리 하여 실험하였다.

- 실험 1: RGB 이미지
- 실험 2: DFT 변환된 이미지
- 실험 3: ELA 필터링한 이미지

실험의 일관성을 위하여 세 모델에 대하여 동일한 하이퍼파라미터를 사용하였으며, 그 종류와 수치는 다음과 같다.

Input shape	512 * 800
Epochs	50
Learning Rate	0.001
Pretrained Weights	ImageNet
Batch Size	64

[표 1] Hyperparameters for training each models.

모델의 종류와, 특징 추출 방법만을 다르게 한 채 실물, 종이 출력, 모니터 출력 신분증 이미지를 분류하도록 각 학습을 설정하였으며, 그에 따른 분류 정확도를 측정하였다. Validation loss가 가장 낮았을 때의 모델 정확도와 학습 중 모든 epoch에서 모델을 저장한 후 최고 정확도를 측정하였다.

	Valid acc	**Test acc (least loss)**	**Test acc (best)**	**특징 추출 방법**
MobileNetV2	99.69	92.30	95.31	Raw Image
MobileNetV2	96.64	81.38	84.92	DFT
MobileNetV2	97.46	82.86	88.9	ELA
EfficientNet_b1	99.82	91.55	96.61	Raw Image
EfficientNet_b1	98.44	85.74	89.21	DFT
EfficientNet_b1	99.13	86.20	90.94	ELA
EfficientNet_b3	99.88	94.58	96.22	Raw Image
EfficientNet_b3	98.49	84.15	87.63	DFT
EfficientNet_b3	99.40	86.10	90.06	ELA

[표 2] Test results

4.2 결과 분석

결과는 모델과 특징 추출 방법에 따라 다르게 나타난다. 전체적으로 모델의 크기가 크고 구조가 복잡할수록 더 높은 성능을 보인다. Raw image로 분류한 모델들은 모두 높은 성능을 보였으며, 이 중 EfficientNet_b3 이 가장 높은 정확도를 보였다.

선행 연구에서 사용한 특징 추출 방법들 중 좋은 성능을 보였던 DFT와 ELA를 사용하여 실험하였다. 신분증 이미지의 디지털적 조작 여부를 판별하는 데 일정한 도움 정도는 될 수 있으나, 선행 연구와는 달리 Raw image만을 사용하여 학습한 것만큼 좋은 성능을 내지 못하는 것을 확인할 수 있다.

DFT, ELA를 사용하였을 때 성능이 잘 나오지 않는 것은 적은 데이터로 인한 fine-tuning의 영향력이 크다고 볼 수 있다. 초기 네트워크의 파라미터로 ImageNet weight을 사용하고 있어 Fine tuning에서는 자연에 가까운 RGB 채널의 입력을 받아야 하는데, DFT, ELA로 처리한 값은 자연과는 멀고, 특히 DFT 같은 경우는 각 RGB 채널별로 변환한 relation이 적은 주파수 이미지를 생성한다. 따라서 pretrained weight를 활용하지 못할 수 있어 더 적은 성능이 나온 것이라 추정하고 있다.

V. 결론 및 향후 연구 방향

본 연구에서는 여러 CNN 모델 중 EfficientNet_b1, EfficientNet_b3, MobileNetV2를 활용하여 이미지 분류 작업을 수행하였다. 또한, raw image, 이산 푸리에 변환(DFT), 그리고 오차 수준 분석(ELA) 세 가지 다른 특징 추출 기법을 사용하였다. 실험 결과, 모든 특징 추출 기법에 대해 EfficientNet_b3 모델이 가장 높은 성능을 보였으며, raw image 기반의 특징 추출에서 가장 나은 성능을 보였다.

하지만 선행 연구와는 달리, DFT와 ELA를 사용했을 때 pretrained weight의 특성상 RGB raw data를 사용하는 것만큼 좋은 성능을 내지 못함을 확인하였다. 특징 추출 방법을 사용하는 것이 우리가 제작한 데이터셋으로 분류하는 것에는 적합하지 않을 수 있음을 보여주며, 새로운 연구의 필요성을 제기한다. 선행 연구와 비슷한 방식으로 제작한 데이터셋을 사용하였음에도 변환하였을 때 성능을 향상시키지 못하는 이유에 대한 후속 연구가 요구된다.

이러한 결과를 바탕으로, 향후 연구 방향 및 실험 계획을 제시하려고 한다. 특징 추출 기법에 따른 모델 최적화 방안에 대한 연구가 필요하다. Pretrain하지 않은 모델을 사용하거나, 기존 모델의 input에 DFT, ELA 처리된 이미지를 추가하는 방안을 생각해 볼 수 있다. 각 특징 추출 기법의 특성에 따라 모델의 weight이나 학습 방식을 최적화함으로써 성능을 개선할 수 있을 것이다. 또한 이 연구에서는 EfficientNet_b1, EfficientNet_b3, 그리고 MobileNetV2 세 가지 모델만을 고려하였다. 그러나 이 미지를 대상으로 하는 다른 CNN 모델 또는 트랜스포머 역시 성능 비교를 위해 고려될 수 있다.

본 연구는 몇 가지 CNN 모델을 사용하여 특징 추출 기법이 처리된 이미지 분류 성능을 비교하였으며, 이를 바탕으로 향후 연구 방향을 제안하였다. 이 연구가 조작된 신분증 이미지 연구에 중요한 시사점을 제공하고, 더 나은 방법론 개발에 기여할 것으로 기대된다.

[참고 문헌]

[1] O'Shea, Keiron, and Ryan Nash. An Introduction to Convolutional Neural Networks. arXiv, 2 Dec. 2015. arXiv.org, https://doi.org/10.48550/arXiv.1511.08458.

[2] Tan, Mingxing, and Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. arXiv, 11 Sept. 2020. arXiv.org, https://doi.org/10.48550/arXiv.1905.11946.

[3] Sandler, Mark, et al. MobileNetV2: Inverted Residuals and Linear Bottlenecks. arXiv, 21 Mar. 2019. arXiv.org, https://doi.org/10.48550/arXiv.1801.04381.

[4] M. Stokkenes, R. Ramachandra and C. Busch, "Biometric Transaction Authentication using Smartphones," 2018 International Conference of the Biometrics Special Interest Group (BIOSIG), Darmstadt, Germany, 2018, pp. 1-5, doi: 10.23919/BIOSIG.2018.8553455.

[5] Shi, Yichun, and Anil K. Jain. DocFace: Matching ID Document Photos to Selfies. arXiv, 6 May 2018. arXiv.org, https://doi.org/10.48550/arXiv.1805.02283.

[6] Shi, Yichun, and Anil K. Jain. DocFace+: ID Document to Selfie Matching. arXiv, 18 Sept. 2018. arXiv.org, https://doi.org/10.48550/arXiv.1809.05620.

[7] P. Perera and V. M. Patel, "Face-Based Multiple User Active Authentication on Mobile Devices," in IEEE Transactions on Information Forensics and Security, vol. 14, no. 5, pp. 1240-1250, May 2019, doi: 10.1109/TIFS.2018.2876748.

[8] Benalcazar, Daniel, et al. Synthetic ID Card Image Generation for Improving Presentation Attack Detection. arXiv, 31 Oct. 2022. arXiv.org, https://doi.org/10.48550/arXiv.2211.00098.

[9] S. Gonzalez, A. Valenzuela and J. Tapia, "Hybrid Two-Stage Architecture for Tampering Detection of Chipless ID Cards," in IEEE Transactions on Biometrics, Behavior, and Identity Science, vol. 3, no. 1, pp. 89-100, Jan. 2021, doi: 10.1109/TBIOM.2020.3024263.