JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 22, No. 2, pp. 339-346, Feb. 2021

Check for updates

심층 합성곱 신경망을 사용한 인스타그램 위조품 판매 게시물 탐지

박 정 은[1] · 김 하 영[2*]
[1]연세대학교 정보대학원 박사과정
[2*]연세대학교 정보대학원 조교수

Detecting counterfeit sales posts on Instagram using deep convolutional neural network

Jeongeun Park[1] · Ha-Young Kim[2*]

[1]Doctor's Course, Graduate School of Information, Yonsei University, Seoul, 03722, Korea

[2*]Assistant Professor, Graduate School of Information, Yonsei University, Seoul, 03722, Korea

[요 약]

2000년대 초부터 SNS가 활발해 지면서 위조품 판매 채널이 SNS로 많이 전이되었는데, 대표적인 온라인 사진 공유 서비스인 인스타그램에서는 2016년 20,892개의 위조품 판매 계정이 있던 것에 비해 2019년 56,769개로 171% 이상 증가하였다. 본 연구는 SNS 환경을 저해하는 위조품 판매 게시물을 탐지하고 더 나아가 위조품 및 가짜 상품의 광범위한 생산을 억제하는 데 도움이 되고자 딥러닝 방법을 사용한 위조탐지 알고리즘을 제안하였다. 연구를 위해 인스타그램에서 총 382,790건의 데이터를 수집하였고, 사람이 파악할 수 없는 이미지 내 특정 패턴까지 분석할 수 있는 심층 합성곱 신경망 분석 방법을 사용하여 위조품 판매 게시물의 특성을 파악하였다. 사전학습 된 심층 합성곱 신경망 모델 5가지를 사용하여 탐지 성능을 비교해 본 결과 가장 성능이 좋은 딥러닝 모델은 이미지만으로도 92%의 성능으로 위조품 판매 게시물을 탐지해 낼 수 있었다. 또한, 경량화 모델 역시 90%의 성능을 내었는데 이를 통해 위조품 판매 게시물 탐지 알고리즘의 모바일 환경에서 실제 사용 가능성을 증명해 보였다.

[Abstract]

This study proposed a detection algorithm using a deep learning method to help detect counterfeit sales posts that hinder the SNS environment and further curb the widespread production of counterfeit. We analyzed 382,790 data collected from Instagram using a deep convolutional neural network (DCNN) and identified the characteristics of counterfeit sales posts. As a result of comparing detection performance using five pre-trained DCNN models, the best performance model was able to detect counterfeit sales posts with 92% performance only with images. In addition, the lightweight model also performed 90% which demonstrates the practical availability of counterfeit sales post detection algorithms in the mobile environment.

색인어 : 위조품 탐지, 딥러닝, 심층 합성곱 신경망, 특징 추출, 인스타그램

Key word : Counterfeit detection, Deep learning, DCNN, Feature extraction, Instagram

http://dx.doi.org/10.9728/dcs.2021.22.2.339

Received 21 December 2020; **Revised** 27 January 2021
Accepted 27 January 2021

***Corresponding Author; Ha-Young Kim**

Tel: +82-2-2123-4294
E-mail: hayoung.kim@yonsei.ac.kr

Ⅰ. 서 론

The Global Brand Counterfeiting Report 2018[1]에 따르면 세계적으로 총 위조품의 양은 2017년 1.2조 달러에 이르렀고, 2020년에는 모든 장비·제품의 위조를 포함하여 1.82조 달러에 이를 것으로 예상하였다. 특히 전 세계 온라인 위조로 인한 손실은 3,300억 달러에 이르는데, 이 중에서도 인터넷을 통한 위조품 판매로 인해 럭셔리 브랜드에서 발생한 손실은 30.3억 달러에 달한다. 이러한 수치에서 볼 수 있듯 위조품의 확산은 이제 더는 오프라인에 한정되지 않고 디지털화되면서 여러 플랫폼을 통해 바이러스처럼 확산되고 있다. 실제로 2019년 4월 Analytics Firm Ghost Data[2]의 조사에 따르면 대표적인 소셜 미디어 플랫폼인 인스타그램의 패션 제품에 관한 모든 게시물의 20%가 위조품이었으며, 활성화된 위조품 판매 계정을 확인한 결과 2016년 20,892개의 계정이 있던 것에 비해 2019년 56,769개로 171% 이상 증가하였다. 다시 말해, 상당수의 위조품 판매자들이 인스타그램을 판매 채널로 이용하고 있다는 것인데 이러한 현상은 인스타그램이 2019년 3월 19일 소매 업체를 위한 In-App 결제 기능을 출시함에 따라 더욱 확산 및 가속될 것이라 예상된다.

이렇듯 사용자들이 제품을 더욱 손쉽게 구매할 수 있는 환경이 조성되면서 더 많은 위조품 판매자들이 인스타그램을 판매 플랫폼으로 활용하게 되었다. 그 결과, 사용자들은 그들의 의지와는 관계없이 위조품 판매 게시물 같은 무수한 광고성, 클릭베이트 성 게시물에 더욱 자주 노출될 수밖에 없게 되었다. 이러한 위조 상품 판매 채널의 다각화 및 온라인화는 사회·경제적 측면에서 더욱 큰 타격으로 이어질 가능성이 있으므로 하루빨리 위조품 판매 근절을 위한 노력이 이루어져야 할 것이다.

학계에서도 위조품과 관련한 연구가 2000년대 초반부터 꾸준히 진행됐는데 대부분의 연구가 위조품 구매 의도에 미치는 영향에 관한 연구[3]-[8]이거나 구매 태도[9]-[11]에 관한 연구로, 이들 연구는 위조품을 구매하는 구매자들에 초점이 맞춰져 있다. 반면에 위조품을 판매하는 사람 혹은 온라인상에서 판매되는 위조품에 관한 연구는 부족한 실정이다. 유사하게, 인터넷상에서 사용자를 저품질 콘텐츠로 유도하는 클릭 베이트 탐지에 관한 연구는[12]-[15] 진행되었다. 이들 연구는 인터넷상에서 사용자들의 호기심을 유발해 저품질 콘텐츠를 포함하고 있는 링크로의 유입을 목적으로 하는 특정 게시물에 대한 탐지를 목표로 하고 있다. 사용자들의 이목을 집중시켜 링크로 유입시키려는 목적 자체는 클릭 베이트나 위조품 판매 게시물이나 동일하다. 하지만 링크로의 유입이 곧 수익으로 이어지는 클릭 베이트와는 다르게 인스타그램에서 위조품을 판매하는 계정은 제품을 판매하고, 홍보하는 것에 더 큰 목표를 둔다. 사용자들의 이목을 집중시키는 요소가 텍스트(클릭 베이트)인지 이미지(위조품 판매 게시물)인지에도 큰 차이가 있다. 따라서 더욱 근본적으로 일반 계정과 위조품을 판매하는 게시물을 판별해내는 연구를 통해 사용자들의 소셜 네트워크 서비스(SNS; social network service)와 같은 온라인 서비스 사용 경험을 저해하는 요소들을 찾아내고, 더 나아가 사회 전체에 부정적 영향을 미칠 수 있는 위조품 판매 계정에 대한 접근을 원천적으로 차단할 수 있는 알고리즘을 제안하는 연구가 이루어져야 한다. 이러한 목적의 일환으로 본 연구는 SNS 환경을 저해하는 위조품 판매 게시물을 탐지하고 더 나아가 위조품 및 가짜 상품의 광범위한 생산을 억제하는 데 도움이 되고자 한다.

본 연구에서는 위조품 판매 게시물을 탐지하기 위해 딥러닝 알고리즘을 이용한다. 특히 딥러닝 알고리즘은 여러 도메인에 적용되며 곳곳에서 새로운 패러다임의 변화를 일으키고 있다. 따라서 본 연구에서는 인스타그램에서 수집한 대량의 데이터를 딥러닝 기술을 사용하여 분석함으로써 사회적 이슈가 되는 위조품 판매 게시물을 탐지하는 방안을 제시하도록 한다.

Ⅱ. 선행연구

2-1 이상 계정 탐지

2004년 페이스북, 2006년 트위터, 2010년 인스타그램 등 2000년대 SNS 서비스가 본격적으로 등장하기 시작하면서 SNS 상에서 이상 계정 탐지에 관한 연구가 활발하게 진행되기 시작했다. 특히 주성분 분석(PCA; principal component analysis) 이나 k-means clustering 방법을 사용한 연구가 많은데[16]-[19], 그중에서 Viswanath et al. 연구진[16]은 일반 사용자의 행동을 정확하게 모델링하고 이로부터 중요한 편차를 이상으로 식별하는 PCA 알고리즘을 적용해 페이스북 네트워크에서 사용자의 유사 활동을 연구하였다. 이와 유사하게, Adewole et al. 연구진[17]은 PCA 및 k-means clustering 알고리즘을 사용하여 스팸 발송자를 탐지하였다. Savyan & Bhanu[18] 는 페이스북에서 사용자들의 반응을 k-means clustering과 코사인 유사도 분석을 사용하여 클러스터링 하였다. 이를 통해 사용자들의 반응 변화를 살펴보고, 비정상적인 동작을 탐지하였다. Lee & Kim[19]은 악성 계정 이름을 가진 스팸 발송자를 그룹화하기 위해 계층적 클러스터링 방식을 제안했다. 그들은 트위터의 유효한 계정 이름으로 마르코프 체인 모델을 훈련하여 일정 패턴을 위반하는 모든 계정 이름은 악성으로 표시하였다.

이러한 연구 이외에도 이상 계정을 탐지하는데 딥러닝 기법을 사용한 연구도 있다. Alom et al. 연구진[20]은 딥러닝 기반의 트위터 스팸 발송자 탐지 알고리즘을 제시했다. 이들은 트윗 텍스트와 사용자의 메타 데이터(계정, 나이, 팔로워/팔로잉 수 등)를 모두 활용해 스팸 발송자를 탐지했다. 비슷하게 Gong et al. 연구진[21]은 위치기반 소셜 네트워크 서비스용 악성 계정 탐지 시스템인 DeepScan을 설계하였다. DeepScan은 딥러닝 기술을 활용해 사용자의 동적 행동을 학습한다. 특히, 사용자 활동의 시계열 분석을 수행하기 위해 장단기 기억 메모리(LSTM; long short term memory)를 사용하였고 이를 통해 0.964의 F1 점수를 달성하였다.

2-2 클릭 베이트 탐지

클릭 베이트란 클릭(click)과 미끼(bait)의 합성어로 사용자들로 하여금 링크를 클릭하고 싶게 유도한 후, 낮은 가치의 정보나 콘텐츠를 제공하는 것을 일컫는다. 머신러닝이나 딥러닝 기술을 활용해 클릭 베이트를 탐지하는 연구는 2016년부터 본격적으로 나오기 시작했다. 가장 먼저 Potthast[22]는 클릭 베이트에 대한 215가지 특성(feature)을 선정한 후, 세 개의 카테고리로 분류하였다. 이러한 특성들을 사용하여 로지스틱 회귀공식(LR; logistic regression), 나이브 베이즈 분류기(NB; Naïve Bayes classifier), 랜덤 포레스트(RF; random forest)로 각각 탐지 모델을 만들어 평가하였다.

Agrawal[23]은 클릭 베이트 탐지에 처음으로 심층 합성곱 신경망을 사용하였다. 레딧, 페이스북, 트위터에서 추출한 데이터를 바탕으로 word2vec으로 임베딩한 텍스트 데이터를 심층 합성곱 신경망의 입력으로 넣고 분류 모델을 만들었다. Zheng et al. 연구진[24]도 이와 비슷한 연구를 수행하였는데, 사전 훈련된 word2vec을 사용해 기사의 헤드라인을 의미론적으로 이해하고 다양한 커널을 사용해 헤드라인의 특성을 찾았다. 이후 심층 합성곱 신경망을 사용하여 분석하였다.

클릭 베이트에 관한 연구는 특히 뉴스의 헤드라인 기사를 사용한 연구가 많다. Geçil[25]은 뉴스의 헤드라인과 부제목을 수집하여 뉴스 기사에서 TF-IDF를 사용하여 식별한 클릭 베이트 헤드라인을 온톨로지 방법을 기반으로 한 텍스트 특징 추출로 클릭 베이트 뉴스의 내용을 보고 요약했다. 이렇게 요약된 버전은 뉴스 기사를 클릭하지 않고도 사용자에게 바로 표시되도록 하였다. Chakraborty et al. 연구진[26]은 wikinews에서 기사 18,513건을 수집하여, 클릭 베이트를 자동으로 감지한 다음 특정 클릭 베이트가 다른 웹 사이트에 나타나지 않도록 차단하는 브라우저 확장 프로그램을 구축하였다. 서포트 벡터 머신(SVM; support vector machine)과 의사 결정 나무(DT; decision tree), RF를 사용하여 클릭 베이트 분류기를 만들고, 차단된 기사의 언어적 패턴을 식별하여 클릭 베이트를 차단하는 시스템을 설계하였다.

Ⅲ. 위조품 판매 게시물 탐지

3-1 데이터 수집 및 전처리

대표적인 소셜 네트워크 서비스인 인스타그램에서 세 단계를 거쳐 데이터를 수집하였다. 첫 번째로 기존 연구[27]와 동일하게 인스타그램 내 상위 인기 태그인 '패션' 카테고리에서 15개의 글로벌 패션 브랜드를 선정한다. 이후, 해당 브랜드명을 해시태그로 사용한 계정 ID를 수집하였다. 마지막으로 각 계정에 대한 구분은 표 1과 같은 기준에 의거하였으며, 인스타그램 사용자 세 명이 각각 계정에 대해 판단한 후, 세 명의 의견이 모두 일치한 계정에 한하여 등록된 이미지 정보를 수집하였다.

최종적으로 일반 계정 100건, 위조품 판매자 계정 100건에 대한 데이터 세트를 구축하였는데, 이때 계정에는 위조품 판매자 계정과 마찬가지로 인스타그램 상에서 상품을 판매하는 일반 판매자 계정이 38건 포함되어 있다. 이는 인스타그램의 생태계를 고려한 것으로, 2017년 기준 인스타그램 사용자는 8억 명이고, 2500만 명의 계정이 비즈니스 계정으로 등록되어 있다[28]. 이는 전체 사용자의 약 3.1%의 비율인데, 정식으로 등록된 비즈니스 계정이 아니더라도 인스타그램 상에서 자신이 판매하는 제품을 업로드하고, 판매 링크 등을 노출하는 일반 판매자 계정이 상당수 존재하는 것을 고려하여 약 6:4의 비율이 되도록 데이터를 수집하였다. 현재 일반 판매자의 경우 그 수가 공식적으로 집계되지 않았기 때문에 정확한 비율을 알 수 없다. 하지만 6:4의 비율은 단순히 제품 사진인지, 인물사진인지에 의해 예측 결과가 결정지어지지 않을 정도로 일반 판매 계정을 포함하되, 인스타그램은 일반 계정이 위주를 이루는 환경이라는 것을 고려하여 결정한 비율이다. 본 연구에서 이미지 수집은 Instalodaer API를 사용하였다.

수집 완료 후, 데이터 전처리는 크게 두 단계로 진행하였다. 첫 번째는 데이터 클렌징 단계로, 한 게시물에 여러 개의 이미지가 포함된 경우 모든 이미지를 수집한 후 1번부터 N 번까지 번호를 부여하였다. 본 연구에서는 위조품 판매 게시물을 탐지하는 것이 목적이므로 위조품 판매자 계정에서 수집된 데이터를 positive로, 일반 사용자 계정에서 수집된 데이터를 negative로 표현하였다. 일부 위조품 판매자 계정에 포함된 위조품과 관련 없는 게시물의 경우 학습 시에는 모델을 robust 하게 만들어주는 데 도움이 될 수 있으나, 예측 시에는 노이즈가 될 수도 있어 테스트 세트에는 포함되지 않도록 사전 점검을 통해 필터링 해주었다.

표 1. 계정에 대한 구분 기준
Table 1. Criteria for classification of counterfeit sales accounts

1	Provides personally accessible information such as phone number, wechat, and whatsapp.
2	Provides information that may be suspicious of the authenticity of a product, such as lowest price, $300, and best quality.
3	Hashtags that are not related to posts are repeatedly used.
4	Posts with the same content are used repeatedly.

표 2. 해시 태그 검색에 사용한 글로벌 패션 브랜드 리스트
Table 2. List of global fashion brands used for hashtag search

#chanel	#louis vuittion	#hermes	#celine	#gucci
#prada	#dior	#rolex	#bottega veneta	#balenciaga
#vetement	#supreme	#max mara	#nike	#adidas

두 번째 단계는 over-sampling이다. Fraud detection이나 anomaly detection과 같이 데이터의 각 클래스가 현저하게 차이 나는 문제를 다룰 때 가장 먼저 고려해야 할 것이 데이터의 불균형문제이다. 해당 연구에서도 클래스별로 각각 100명의 계정에 등록된 이미지 정보를 수집하였는데 negative 데이터 세트가 positive 데이터 세트의 4.4 배 정도로 각 클래스의 불균형문제가 두드러졌다. 딥러닝 알고리즘은 각 클래스의 개수가 비슷할 때 좋은 결과를 내는데, 이처럼 하나의 클래스 개수가 다른 클래스에 비해 많게 되면 알고리즘은 데이터 세트가 적은 positive 클래스를 negative 클래스로 오분류 하는 경향이 나타날 것이다. 따라서 이러한 데이터 불균형문제를 해결하기 위해 본 연구에서는 over-sampling 방법의 하나인 repetition을 사용하여 학습 데이터 세트의 minority 클래스를 추가해 주었다. 데이터 전처리를 거쳐 최종적으로 분류된 데이터 세트는 총 382,790건이다.

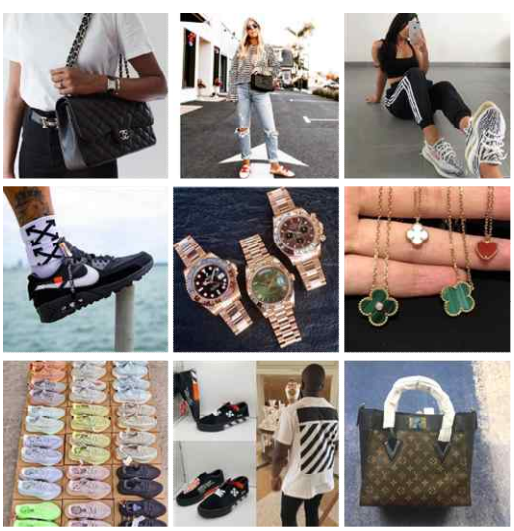

그림 1. Positive 이미지 데이터 세트
Fig. 1. Positive image dataset

그림 2. Negative 이미지 데이터 세트
Fig. 2. Negative image dataset

분석에 앞서, 그림 1, 2와 같이 수집된 데이터를 살펴보면, positive 데이터는 해당 상품을 착용한 일반인들의 사진이나 오리지널 브랜드에서 자신들의 인스타그램에 게시한 사진을 카피하여 무단으로 사용한 경우를 종종 찾아볼 수 있었다. 이 경우, negative 데이터 세트와 사진만으로는 구분이 어려울 수 있으나, 대체로 원본 사진이 아니므로 사진의 해상도가 낮은 경우가 많았다. 또한, positive 데이터의 경우 조잡하게 찍힌 사진이 많았고, 제품과 관련된 사진이 대다수를 이뤘다. 동일한 사진이 여러 차례 등장하기도 했다. 반면에 negative 데이터의 경우 사진의 카테고리가 훨씬 다양했다. 인물이나 풍경 사진도 존재하였고, 육안으로 보이는 사진의 퀄리티가 positive 데이터의 퀄리티 보다 훨씬 좋았다. 같은 사진이 반복되는 경우는 드물었다. 이러한 특징들 이외에 딥러닝 모델은 사람 눈에는 보이지 않지만, 두 클래스 간에 존재하는 패턴들을 학습할 수 있을 것이기 때문에 본 연구에서는 아래와 같은 분석 방법을 사용한다.

3-2 분석 방법

본 연구에서는 인스타그램 게시물이 포함하고 있는 무수한 정보 중 이미지 데이터만 가지고 위조품 판매 게시물을 탐지한다. 인스타그램은 기본적으로 시각적 자료 (이미지, 동영상)가 주가 되는 플랫폼이다. 특히 인스타그램은 개인이 이미 팔로우하는 계정과 유사한 계정인 시드 계정에서 탐색 피드에 콘텐츠를 소싱해 주고 있다. 이때 소싱되는 콘텐츠는 이미지와 동영상 뿐이며, 이미지를 클릭해야지만 사용자가 게재한 텍스트까지 확인할 수 있다. 기본적으로 클릭 베이트와 유사하게 사용자들의 이목을 끄는 것이 목적인 위조품 판매 계정의 특성을 고려했을 때, 이미지에 가장 많은 정보가 들어 있으리라 판단하여 이미지만을 가지고 위조품 판매 게시물을 탐지하는 모델을 설계하였다.

위조품 판매 계정이 업로드한 데이터와 다른 일반적인 계정에서 업로드한 데이터 간에 어떠한 차이가 있는지 알아보기 위해서 심층 합성곱 신경망으로 일컫는 딥러닝 방법을 사용하고자 한다. 심층 합성곱 신경망을 사용하여 이미지 데이터를 분석하면 이미지가 가지고 있는 공간 정보를 유지하면서 사람이 파악할 수 없는 이미지 내 특정 패턴까지 분석할 수 있다는 장점이 있다.

본 연구에서는 이미지넷 데이터 세트에서 사전 학습된 Inception_v3, VGG16, ResNet50_v2, EfficietNetB0, MobileNet_v2[29]-[33] 모델을 Transfer-learning한 후 연구 문제에 맞게 Fine-tuning 하였다. 백본 모델의 선정 이유는 다음과 같다. 2014년 ILSVRC(imagenet large scale visual recognition challenge)에서 각각 1위와 2위를 차지한 GoogLeNet과 VGGNet은 높은 성능으로 ILSVRC에서 우승을 차지한 모델들이다. 특히 VGGNet은 ILSVRC 에서는 2위를 차지한 모델이지만 굉장히 간단한 구조로 이루어져, 복잡한 형태의 GoogLeNet보다 딥러닝 연구자들에게 더 인기를 끌었다. 이러한 좋은 성능과 간단한 구조라는 이유로 본 연구에서는 Inception_v3와 VGG16 모델을 적용하였다.

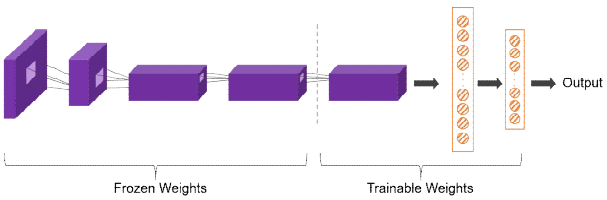

그림 3. 전체 프레임워크
Fig. 3. Overall framework

ResNet은 Microsoft Research에서 제안한 구조이며, 2015년 ILSVRC 대회에서 우승을 차지한 모델로, Top -5 error가 겨우 3.6%로 사람의 분류 수준인 5% 내외를 뛰어넘었다. ResNet 이란 이름은 해당 논문에서 제안한 핵심 아이디어인 residual block에서 유래했으며, 실제로 이 residual block 하나만 알면 전체 모델을 이해할 수 있을 정도로 단순하면서도 효과적인 구조로 되어 있다.

다음으로 분류 모델 중 State-of-the-Art(SOTA) 알고리즘인 EfficientNet을 사용한다. 구글 연구팀에 의해 제안된 EfficientNet은 위의 두 모델에 비해 훨씬 적은 연산량과 파라미터를 가지면서도 높은 성능을 자랑한다. 심층 합성곱 신경망 구조를 효과적으로 스케일 업하는 방법을 통해 모델의 효율성 및 정확도를 극대화시킨 EfficientNet 중에서도 가장 연산량 및 파라미터수가 적은 B0모델을 적용한다.

마지막으로 MobileNet은 본 연구의 향후 사용 가능성을 고려하여 선정하였다. 인스타그램은 주로 모바일 환경에서 사용하기 때문에 일반적인 심층 합성곱 신경망 모델들을 사용하게 되면 경량화를 위한 작업이 추가로 필요할 수도 있다. 하지만 MobileNet은 기존의 convolution layers를 depth-wise separable convolution으로 대체하여 모바일과 같이 컴퓨터 파워가 제한적인 상황에서 사용할 수 있도록 설계한 경량화 모델이다. 따라서 본 연구에서는 위조품 판매 게시물 탐지 알고리즘의 실사용성을 고려한 MobileNet을 마지막 백본 후보로 선정하였다.

Ⅳ. 실험 및 결과

4-1 실험 방법

실험에 앞서, 이미지 데이터의 사이즈를 224 × 224 사이즈로 맞춰주었다. 인스타그램의 경우 서비스 초기에 정사각형 이미지만 업로드 가능하였지만, 이후 몇 번의 업데이트를 거쳐 다양한 사이즈의 이미지를 업로드 할 수 있게 되었다. 수집한 데이터 세트의 인풋 사이즈가 매우 다양하여 224 × 224 크기로 크롭하거나 비율을 고려하지 않고 축소하는 방법은 데이터를 훼손시킬 수 있다고 판단하여, 긴 축에 맞춰 같은 크기를 가지도록 사진의 나머지 부분에 0패딩을 준 다음 224 × 224 크기로 축소하는 방법을 선택하였다.

전체 이미지 데이터는 382,790건으로 학습을 위해 7:1:2의 비율로 학습/검증/테스트 세트로 나눠주었다. 그 결과, 학습 데이터는 negative 217,203건, positive 49,069건이며, 검증 데이터는 negative 32,949건, positive 6,388건, 테스트 데이터는 negative 63,180건, positive 14,001건으로 분류되었다. 이때 학습 데이터 세트의 positive 클래스는 over-sampling을 해주었기 때문에 실제로는 245,345건이 학습에 사용되었다. 또한, 모델의 성능을 향상시키기 위해 학습 데이터 수를 augmentation을 사용하여 더욱 늘려주었다. 데이터 augmentation은 이미지를 사용할 때마다 임으로 변형을 가함으로써, 오버피팅을 방지해주는데 즉, 모델이 학습 데이터에만 맞춰지는 것을 방지하고 새로운 이미지도 잘 분류할 수 있도록 해준다. 데이터 augmentation을 위해 keras에서 제공하는 ImageDataGenerator 클래스를 사용하였다. ImageDataGenerator는 전체 데이터를 메모리에 올려서 작업하는 것과는 달리 model fitting process를 통해서 Just-In-Time으로 이미지를 증식시킨다.

각 모델은 이미지넷 데이터에서 사전 학습되었으며, 기존 모형의 상단 층은 제거한 후, global average pooling과 모델에 따라 적절한 dense 층 및 prediction 층을 더해주었다. fine-tuning을 할 때, prediction 층과 가까운 일부 convolution 층을 제외한 모든 base 층은 동결시켜주었다. 배치 사이즈는 64로 주었으며, 초기 학습률이 1e-3인 Adam optimizer을 사용하였다. 또한, 학습률은 스케줄링을 통해 검증 정확도 변동을 기반으로 $\sqrt{0.1}$씩 5 epoch patience를 가지고 줄어들 수 있도록 하였다.

본 연구는 위조품 판매 게시물인지 아닌지를 구분하는 이진 분류의 문제이다. 따라서 모델의 마지막 층의 활성 함수는 수식 1과 같이 시그모이드 활성 함수를 사용하였다. 손실 함수는 수식 2와 같이 크로스 엔트로피가 최소가 되는 방향으로 학습되게 한다.

$$s(z) = \frac{1}{1+e^{-z}} \tag{1}$$

$$BCE(x) = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(h(x_i;\theta)) + (1-y_i)\log(1-h(x_i;\theta)) \tag{2}$$

전체적인 모델의 학습 프로세스는 그림 4와 같다. 계정 구분 기준에 맞는 200개의 계정에서 수집한 데이터의 클래스 비율을 맞춰준 후, 성능 향상을 위해 데이터 augmentation을 진행한다. 이후 사전 학습된 모델 5가지를 연구 문제에 맞게 fine-tuning하고 성능을 평가한다.

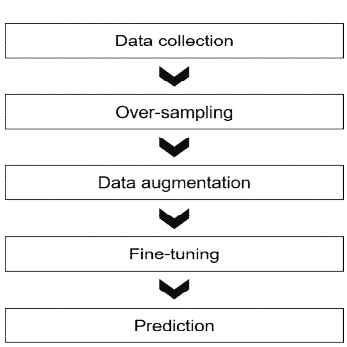

그림 4. 모델 학습 프로세스
Fig. 4. Model training process

4-2 성능 평가 및 분석

총 5가지의 백본 모델을 fine-tuning 하여 위조품 판매 게시물을 탐지해 보았다. 본 연구는 앞서 설명한 것과 같이 데이터에 불균형문제가 존재한다. 학습 시에 over-sampling을 통해 각 클래스의 밸런스를 조절해 주었지만, 실제 inference 단에서는 데이터 클래스를 조절해 주지 않았기 때문에 모델의 정확한 성능을 알아보기 위해 F1 점수를 사용하였다. F1 점수를 통해 클래스별 모델의 예측 정확도를 살펴본 결과는 표 3과 같다. 가장 성능이 좋은 모델은 Inception_v3로 F1 점수의 값이 0.9201이다. 다음으로 ResNet50_v2, EfficientNetB0, MobileNet_v2, VGG16 순으로 나타났는데, 이 4가지 모델은 VGG16을 제외하면 F1 점수에 큰 차이가 나타나진 않는다.

실험 결과를 통해 다음과 같은 사실을 알 수 있었다. 첫째, VGG16은 본 연구의 백본 모델로 적합하지 않다. VGG16은 이름에서도 알 수 있듯 16개의 층을 쌓아 만든 모형이다. 일반적으로 딥러닝 구조는 층이 깊을수록, 층마다 뉴런 수가 많을수록 high-level feature를 학습할 수 있으므로 데이터가 충분하다면 성능을 향상시킬 수 있다. 이러한 사실에 비춰 보았을 때, 총 14,846,530개의 파라미터를 가지는 VGG16의 결과가 다른 모델들에 비하여 좋지 않은 이유는 16개의 층으로는 해당 문제를 푸는 데 충분한 high-level feature를 학습하지 못했기 때문으로 볼 수 있다.

둘째, 경량화 모델인 MobileNet_v2의 결과는 위조품 판매 게시물 탐지 모델이 실제 환경에서 사용될 가능성을 보여주었다. 본 연구에 사용된 MobileNet_v2는 파라미터 수가 2,914,882로 23,903,010의 파라미터를 가진 Inception_v3에 비하면 약 8.2배 적다. 파라미터의 수는 적지만, 파라미터가 가지는 표현력을 가능한 한 유지하면서 불필요한 가중치를 최대한 없앴기 때문에 다른 딥러닝 모델들에 못지않은 성능을 보여주었다고 할 수 있다. 실제 모바일 환경을 고려한다면, MobileNet_v2을 사용하여 충분히 위조품 게시물을 탐지할 수 있을 것으로 생각한다. 또한, 이러한 결과를 통해 일반적으로 딥러닝 모델들이 과파라미터화(over-parameterization)화 되어있다는 사실도 확인할 수 있었다.

셋째, 모델의 1종 오류와 2종 오류, 즉, 실제 negative인 정답을 positive라고 예측한 것과 실제 positive인 정답을 negative라고 예측한 것의 비율을 살펴보면 가장 성능이 좋은 Inception_v3와 MobileNet_v2 간에 큰 차이가 없었다. Inception_v3의 1종 오류의 개수는 1583개로 negative 데이터의 2.5%를 잘못 예측하였다. MobileNet_v2의 경우는, 1종 오류의 개수는 2151개로 3.4%를 잘못 예측하였다. 2종 오류는 Inception_v3가 721개로 positive 데이터의 5.1%를 잘못 예측하였고, MobileNet_v2의 2종 오류는 713개로 5.0%를 잘못 예측하였다. 이러한 결과에서 알 수 있듯, 두 모델 모두 실제 위조품 판매 계정을 분류해 내는 목적에 따르면 성능의 차이가 거의 없다고 할 수 있다. 해당 모델들이 실제 위조품 판매 계정 탐지에 사용된다면 일반 계정을 위조품 판매 계정이라고 오분류 할 경우, 사용자들의 불편을 초래할 수 있다. 하지만 Inception_v3나 MobileNet_v2 모두 1종 오류율이 2종 오류율보다 작게 나타났다. 2종 오류의 경우, positive 데이터를 추가 수집하여 positive에 대한 정보를 모델이 더욱 많이 학습한다면 줄어들 수 있을 것이다.

표 3. 딥러닝 모델별 위조품 판매 게시물 탐지 성능
Table 3. Detection performance of counterfeit sales posts by deep learning model

Model	Precision	Recall	F1 Score
Inception_v3	0.8935	0.9485	0.9201
VGG16	0.8358	0.9058	0.8694
ResNet50_v2	0.8773	0.9393	0.9072
EfficientNetB0	0.8749	0.9333	0.9031
MobileNet_v2	0.8607	0.9491	0.9027

Ⅴ. 결 론

본 연구는 대표적인 SNS인 인스타그램에서 사용자들의 서비스 사용 경험을 저해하고, 사회적 악영향을 미치는 위조품 판매 게시물을 탐지하기 위한 딥러닝 모델을 제안하였다. 다양한 종류의 위조품이 존재하지만, 특히 패션 브랜드를 위조한 제품은 SNS를 통해 빠르게 확산되며 사회적·경제적으로 많은 문제를 야기하고 있다. 하지만 이와 관련된 연구는 아직 미비한 실정으로, 이러한 사회적 문제를 해결하기 위한 목적으로 본 연구를 진행하였다.

연구를 위해 인스타그램내 위조품 판매 계정 100개를 포함한 총 200개의 계정에서 382,790건의 이미지 데이터를 수집하였다. 수집된 데이터를 분석하기 위해 여러 분류 문제에서 사용되는 Inception, VGG, ResNet 모델과 SOTA 모델인 EfficientNet, 경량화 모델인 MobileNet을 사용하였다. 탐지 정확도를 비교해 본 결과, Inception 모델이 가장 좋은 성능을 보여주었다. Inception 모델 다음으로, VGG를 제외한 ResNet과 EfficientNet, MobileNet 모두 비슷한 성능을 보여주었는데 이를 통해 경량화 모델인 MobileNet을 사용한 실제 모바일 환경에서의 사용 가능성도 확인해 보았다.

본 연구에서는 이미지나 비디오 데이터가 주가 되는 인스타그램의 특성을 고려하여, 이미지 데이터만을 사용하여 위조품 판매 게시물을 탐지해 내는 모델을 제안하였다. 하지만 향후 텍스트나 좋아요 수, 팔로잉/팔로워 수 등 더 다양한 feature들을 사용하여 모델의 성능을 향상한다면 실제 탐지 모델로 충분히 사용할 수 있을 것으로 예상한다.

감사의 글

본 연구는 2020년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(NRF-2020R1F1A1071527)에 의하여 이루어진 연구로서, 관계부처에 감사드립니다.

참고문헌

[1] Global Brand Counterfeiting Report, 2018, R Strategic Global, 2017.

[2] A. Stroppa, D. Gatto, L. Pasha, B. Parrella, Instagram and counterfeiting in 2019: new features, old problems, Analytics Firm Ghost Data, 2019.

[3] Zhang, Y. Fei, Bae, S. W. Kim, W. Min, "An Empirical Research of Purchasing Intentions toward Imitations in Korean and Chinese Customers," The Journal of North-East Asian Cultures, Vol. 53, pp. 33-46, December 2017.

[4] C. Shuge, M. J. Kim, "Effects of China Online Market Counterfeit Products Message on Purchase Intention," Journal of the Korea Contents Association, Vol. 18, No. 2, pp. 81-91, February 2018.

[5] S. Y. Yu, "Causal Relation Analysis of the Motivation and Benefits Factors Affecting Customers' Purchase Intention of Counterfeit Goods," Journal of Digital Convergence, Vol. 10, No. 11, 287-293, 2012.

[6] I. S. Jung, "A Counterfeit Goods use Culture in China Consumer : focus on Motivation, Satisfaction and Dissatisfaction Factor," Journal of Digital Convergence, Vol. 10, No. 10, pp. 177-185, 2012.

[7] K. Wilcox, H. M. Kim, & S. Sen, "Why do consumers buy counterfeit luxury brands?," Journal of marketing research, Vol. 46, No. 2, pp. 247-259, 2009.

[8] V. Cordell, N. Wongtada, & R. L. Kieschnick Jr, "Counterfeit purchase intentions: role of lawfulness attitudes and product traits as determinants," Journal of Business Research, Vol. 35, No. 1, pp. 41-53, 1996.

[9] J. J. Yu, H. K. Goo, "The Effects of Social-Face Sensitivity and Conspicuous Consumption on Attitude of Luxury Goods and Counterfeits - Focusing on a Comparative Study of Korea and China -," Review of Industry and Management, Vol. 31, No.1, pp. 21-41, June, 2018.

[10] K. S. Kim, "A Comparative Study on Korea and China consumer of counterfeit attitudes and satisfaction and dissatisfaction factors," Journal of Digital Convergence, vol.11, no.5, pp. 169-178, 2013.

[11] C. S. Hwang, "High School Students' Buying Attitudes toward Counterfeit Jeans Relative to Their Self-Concept," Journal of the Korean Home Economics Association, Vol. 46, No. 5, pp.9-17, 2008.

[12] M. Potthast, S. Köpsel, B. Stein, & M. Hagen, "Clickbait detection," In European Conference on Information Retrieval, 2016.

[13] A. Agrawal, "Clickbait detection using deep learning," In 2016 2nd International Conference on Next Generation Computing Technologies (NGCT) IEEE, pp. 268-272. October, 2016.

[14] A. Geçkil, A. A. Müngen, E. Gündogan, & M. Kaya, "A clickbait detection method on news sites," In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, pp. 932-937, August, 2018.

[15] A. Chakraborty, B. Paranjape, S. Kakarla, & N. Ganguly, "Stop clickbait: Detecting and preventing clickbaits in online news media," In 2016 ieee/acm international conference on advances in social networks analysis and mining (asonam), IEEE, pp. 9-16, August, 2016.

[16] B. Viswanath, M. A. Bashir, M. Crovella, S. Guha, K. P. Gummadi, B. Krishnamurthy, A. Mislove, "Towards detecting anomalous user behavior in online social networks," In Proceedings of the 23rd USENIX Security Symposium (USENIX Security), 2014.

[17] K. S. Adewole, T. Han, W. Wu, H. Song, & A. K. Sangaiah, "Twitter spam account detection based on clustering and classification methods," The Journal of Supercomputing, Vol. 76, No. 7, pp. 4802-4837, 2020.

[18] P. V. Savyan, & S. M. S. Bhanu, "Behaviour profiling of reactions in facebook posts for anomaly detection," In 2017 Ninth International Conference on Advanced Computing (ICoAC,) pp. 220-226, 2017.

[19] S. Lee, & J. Kim, "Early filtering of ephemeral malicious accounts on Twitter," Computer Communications, Vol. 54, pp. 48-57, 2014.

[20] Z. Alom, B. Carminati, & E. Ferrari, "A deep learning model for Twitter spam detection," Online Social Networks and Media, 18, 100079, 2020.

[21] Q. Gong, Y. Chen, X. He, Z. Zhuang, T. Wang, H. Huang, & X. Fu, "DeepScan: Exploiting deep learning for malicious account detection in location-based social networks," IEEE Communications Magazine, Vo. 56, No. 11, pp. 21-27, 2018.

[22] M. Potthast, S. Köpsel, B. Stein, & M. Hagen, "Clickbait detection," In European Conference on Information Retrieval pp. 810-817, March 2016.

[23] A. Agrawal, "Clickbait detection using deep learning," In 2016 2nd International Conference on Next Generation Computing Technologies (NGCT), pp. 268-272, October 2016.

[24] H. T. Zheng, J. Y. Chen, X. Yao, A. K. Sangaiah, Y. Jiang, & C. Z. Zhao, “Clickbait convolutional neural network,” Symmetry, Vol. 10, No. 5, pp. 138, 2018
[25] A. Geçkil, A. A. Müngen, E. Gündogan, & M. Kaya, “A clickbait detection method on news sites,” In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) pp. 932-937, August 2018.
[26] A. Chakraborty, B. Paranjape, S. Kakarla, & N. Ganguly, “Stop clickbait: Detecting and preventing clickbaits in online news media,” In 2016 ieee/acm international conference on advances in social networks analysis and mining (asonam) pp. 9-16, August 2016.
[27] Y. I. Ha, & M. Y. Cha, “A Model for Detecting Online Clickbait Spams,” The Korean Institute of Information Scientists and Engineers, pp. 279-281, June 2017.
[28] Statista. Number of active Instagram business profiles from September 2016 to November 2017 [Internet]. Available: https://www.statista.com/statistics/222243/number-of-instagram-business-acounts/
[29] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, & A. Rabinovich, “Going deeper with convolutions,” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1-9, 2015.
[30] K. Simonyan, & A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014
[31] K. He, X. Zhang, S. Ren, & J. Sun, “Deep residual learning for image recognition,” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778, 2016.
[32] M. Tan, & Q. V. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” arXiv preprint arXiv:1905.11946. 2019.
[33] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, & H. Adam, “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” arXiv preprint arXiv:1704.04861. 2017.

박정은(Jeongeun Park)

2012년 : 연세대학교 전기전자공학과 (학사)
2020년 : 연세대학교 정보대학원 UX (석사)
2020년～현 재: 연세대학교 정보대학원
비즈니스 빅데이터 분석 박사과정

2012년～2014년: KT 융합기술원
2014년～2017년: ㈜ 레어마켓
※관심분야 : 머신러닝(Machine Learning), 딥러닝(Deep Learning), XAI(explainable AI), UX(User Experience)

김하영(Ha-Young Kim)

2000년 : 경희대학교 수학과 (학사)
2007년 : 퍼듀대학교 수학과 및 계산금융 (석사)
2010년 : 퍼듀대학교 수학과 (박사)

2011년～2016년: 삼성전자 종합기술원
2016년～2019년: 아주대학교 금융공학과 교수
2019년～현 재: 연세대학교 정보대학원 비즈니스 빅데이터 분석 교수
※관심분야 : 머신러닝(Machine Learning), 딥러닝(Deep Learning), 수학 및 계산금융(Computational and Mathematical Finance), 확률 과정(Stochastic Process), 확률 이론(Probability Theory)