Journal of Korea Multimedia Society Vol. 22, No. 12, December 2019(pp. 1447-1456)
https://doi.org/10.9717/kmms.2019.22.12.1447

합성곱신경망 기반의 StyleGAN 이미지 탐지모델

김지연[†], 홍승아[††], 김하민[†††]

A StyleGAN Image Detection Model Based on Convolutional Neural Network

Jiyeon Kim[†], Seung-Ah Hong[††], Hamin Kim[†††]

ABSTRACT

As artificial intelligence technology is actively used in image processing, it is possible to generate high-quality fake images based on deep learning. Fake images generated using GAN(Generative Adversarial Network), one of unsupervised learning algorithms, have reached levels that are hard to discriminate from the naked eye. Detecting these fake images is required as they can be abused for crimes such as illegal content production, identity fraud and defamation. In this paper, we develop a deep-learning model based on CNN(Convolutional Neural Network) for the detection of StyleGAN fake images. StyleGAN is one of GAN algorithms and has an excellent performance in generating face images. We experiment with 48 number of experimental scenarios developed by combining parameters of the proposed model. We train and test each scenario with 300,000 number of real and fake face images in order to present a model parameter that improves performance in the detection of fake faces.

Key words: Deep Learning, Generative Adversarial Network, Convolutional Neural Network, Fake Image Detection, Face Detection

1. 서 론

딥러닝 기반의 이미지 프로세싱 기술이 발전하면서 진위 여부가 어려울 정도의 고품질 미디어 콘텐츠가 인공지능에 의해 생성되고 있다. 10년 전만 하더라도 가짜 이미지나 합성사진을 생성하기 위해서 Adobe Photoshop, Illustrator 등의 도구를 사용해야 했지만, 최근에는 이미지 편집 도구를 사용하지 않고도 딥러닝을 통해 정교한 가짜 이미지를 생성하는 것이 가능해졌다. 이를 가능하게 하는 대표적인 딥러닝 모델은 GAN(Generative Adversarial Network)[1]이다. 이미 영상·이미지·텍스트 생성 등 다양한 분야에 GAN이 활발히 사용되고 있으며, 2019년 2월에는 NVIDIA에서도 GAN 기반으로 유명 연예인의 얼굴 이미지를 학습한 후, 가짜 얼굴 이미지를 생성하는 기술을 발표하였다[2]. GAN은 사람 얼굴뿐만 아니라, 동물, 풍경 등 어떤 객체라도 인공지능이 현실과 거의 유사하게 이미지를 만들어 낼 수 있게 하는 진화된 이미지 처리기술로 활용되고 있다. 그러나 딥러닝에 의해 가짜 미디어 생성이 쉬워질수록, 이러한 기술이 정치적 또는 상업적으로 악용될 수 있는 소지 또한 많아졌다. 올해 9월에는 트럼프 미국 대통령이

※ Corresponding Author : Jiyeon Kim, Address: (01797) Hwarang-ro 621, Nowon-gu, Seoul, South Korea, TEL : +82-10-6282-3886, FAX : +82-50-4286-3886, E-mail : jykim07@swu.ac.kr
Receipt date : Nov. 28, 2019, Revision date : Dec. 12, 2019
Approval date : Dec. 16, 2019
[†] Center for Software Educational Innovation, Seoul Women's University
[††] Dept. of Information Security, Seoul Women's University (E-mail : ghdtmddk1516@swu.ac.kr)
[†††] Dept. of Information Security, Seoul Women's University (E-mail : hamin7022@swu.ac.kr)
※ This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (NRF-2018R1D1A1B07050543)

백악관에서 우스꽝스러운 연설을 하는 가짜 영상이 유포되어 내년 미국 대선을 앞두고 딥페이크(deep-fake)에 대한 정치계의 우려가 커지고 있다. 또한, 페이스북 CEO인 마크 저커버그가 본인이 전 세계 사람들의 삶과 데이터를 훔쳐 조종하고 있다고 인터뷰하는 가짜 영상이 유포되어 페이스북 사용자들의 이탈을 종용하는 상황이 일어나기도 했다. 이러한 딥페이크가 명예훼손, 신분 도용, 음란물 제작, 사생활 침해 등의 범죄에 악용된다면 사회적 혼란이 가중될 것이다. 따라서 육안으로는 판별이 어려운 고품질의 딥페이크를 탐지하기 위한 연구가 필요하다. DCGAN (Deep Convolutional GAN)[1], BEGAN(Boundary Equilibrium GAN)[2] 등 기존에 발표된 GAN 모델들을 활용한 가짜 이미지 탐지 연구들은 다수 존재하지만, 최근에 발표된 StyleGAN 이미지 탐지 연구는 아직 미비한 실정이다. StyleGAN은 얼굴 이미지를 생성하는 데에 최적화된 모델로서 고품질의 가짜 이미지 생성이 가능하기 때문에 범죄에 악용될 소지 또한 높다. 따라서 본 논문에서는 NVIDIA에서 개발한 StyleGAN 이미지를 탐지하기 위한 모델을 대표적인 이미지 데이터 딥러닝 모델인 합성곱신경망(Convolutional Neural Network, CNN) 기반으로 제안한다. 제안된 모델은 여러 모델 파라미터를 가지도록 설계되며, 파라미터의 조합으로 48개의 실험 시나리오를 생성하여 진짜 얼굴 데이터 15만 개와 StyleGAN으로 생성된 얼굴 데이터 15만 개를 활용하여 학습 및 평가한다. 또한, 잘못 탐지된 결과에 대한 상세분석을 실시함으로써 제안된 모델의 진짜 및 가짜 이미지 탐지 율에 영향을 미치는 모델 파라미터를 제시한다.

본 논문의 구성은 다음과 같다. 2장에서 본 논문의 기본 이론이 되는 GAN 및 StyleGAN에 대해 설명하고, GAN을 활용한 가짜 이미지 탐지 연구 동향을 조사한다. 3장에서는 본 연구에서 제안하는 StyleGAN 얼굴 이미지 탐지 모델을 CNN 기반으로 설계하고, 4장에서는 여러 실험 시나리오를 개발하여 제안된 모델을 학습 및 평가한다. 마지막으로 5장에서는 결론 및 향후 연구를 제시한다.

2. 이 론

2.1 GAN (Generative Adversarial Networks)

생성적 적대 신경망이라고 불리는 GAN은 2014년 Ian[3]에 의해 제안된 심층 신경망으로서 비지도 학습모델에 속한다. GAN은 이미지를 만드는 생성기(generator) 모델과 이미지가 진짜인지 가짜인지 구별하는 감별기(discriminator) 모델로 구성된다. 생성기 모델은 특정한 분포로 데이터를 생성하여 이 데이터를 원본 데이터와 함께 감별기 모델에 제공하고, 감별기 모델은 제공된 데이터가 원본 데이터인지 생성기에서 생성한 데이터인지를 감별한다. 이렇게 경쟁적으로 생성과 감별이 반복되면서 감별기 모델도 진짜인지 가짜인지 판단하기 힘든 이미지를 생성기 모델이 생성해 낼 수 있게 된다.

2.2 StyleGAN(Style Generative Adversarial Network)

StyleGAN[4]은 2018년 12월 NVIDIA 연구원이 개발하여 2019년 2월에 공개된 새로운 GAN 모델이다. GAN이 발표된 이후, DCGAN, BEGAN, PGGAN (Progressive Growing of GAN)[5] 등 기존의 GAN 모델을 보완하는 모델들이 다수 제안되었다. 그러나 이러한 모델이 사람 얼굴 이미지 생성에 활용될 경우, 부자연스러운 이미지가 생성되거나 이미지의 성별, 연령 등 세부적인 사항을 조절하기가 매우 어렵다는 한계가 있었다. NVIDIA에서는 이러한 단점들을 보완하여 사람 얼굴 이미지 생성에서 높은 품질을 보이는 StyleGAN을 공개하였다. StyleGAN은 이미지를 스타일(style)의 조합으로 보고, 생성기 모델의 각 계층에 스타일 정보를 추가하는 방식으로 이미지를 생성한다. 스타일 정보는 이미지의 성별부터 머리색상, 피부 색상까지 포함한다. 이를 통해 StyleGan은 기존 GAN 모델들보다 더욱 안정적이고 향상된 품질의 얼굴 이미지를 생성할 수 있다.

2.3 GAN을 활용한 가짜 이미지 탐지 연구 동향

본 논문의 모델을 제안하기에 앞서 인공지능 기반의 가짜 이미지를 탐지하는 관련 연구들의 동향을 살펴보고자 한다. 가짜 이미지 탐지 연구들은 GAN으로 생성한 이미지를 딥러닝이나 머신러닝 기반으로 탐지하는 연구, 그리고 위조 이미지를 탐지하기 위해 GAN을 활용하는 연구 등이 존재한다. 자세하게 살펴보면 DCGAN, BEGAN, PGAN, WGAN (Wasserstein GAN), WGAN-GP(WGAN with Gradient Penalty) 등 다중의 GAN으로 생성한 얼굴 이미지를 CNN 기반으로 탐지하는 연구 [6-8], DCGAN

을 포함하여 총 5종류의 GAN 이미지를 수집하여 머신러닝 알고리즘인 RFC(Random Forest Classifier), SVM(Support Vector Machine), LC(Linear Classifier)로 탐지하는 연구[9], 그리고 여러 종류의 GAN 이미지의 성능을 탐지 정확도 기반으로 비교하는 연구[10]가 있다. 또한, 이미지 위변조 탐지를 위해 GAN을 활용하는 연구로는 포렌식 얼굴 위변조 탐지를 위해 DCGAN 및 PGGAN으로 얼굴 이미지를 생성하고 VGG-net 기반으로 탐지한 연구[11], 위성 이미지 위변조 탐지를 위해 오토인코더(auto-encoder)와 one-class SVM으로 모델을 생성하는 연구[12]가 수행되었다. 이 밖에도 딥러닝과 동시 발생 메트릭을 활용하여 CycleGAN[13], StarGAN[14] 이미지를 탐지하는 연구[15], WGAN-GP, DCGAN, PGGAN 이미지에 사전 영상 처리를 한 후 포렌식 CNN 모델이 이미지를 분류하는 연구[16] 등이 존재한다. 얼굴 이미지를 다루는 많은 연구들이 진짜 얼굴 이미지로서 CelebA 데이터셋을 사용하고, 가짜 이미지는 DCGAN, WGAN 등 기존에 발표된 다양한 모델들을 활용하여 생성하고 있지만, 최근에 발표된 StyleGAN을 활용한 연구들은 미비한 실정이다. 이에 본 연구에서는 StyleGAN으로 생성된 가짜 이미지를 탐지하기 위한 모델을 CNN 기반으로 제안하고자 한다.

3. 제안한 방법

StyleGAN으로 생성된 가짜 얼굴 이미지 탐지를 위해서는 StyleGAN 이미지 뿐 아니라, 진짜 얼굴 이미지 데이터 수집이 필요하다. 본 논문에서는 진짜 얼굴 이미지 데이터셋으로서 96×96(dpi) 해상도를 가진 CelebA[6] 데이터 총 202,599장 중 15만 장을 사용하였고, 가짜 얼굴 이미지 데이터셋으로는 NVIDIA에서 공개한 styleGAN 이미지 15만 장을 사용하였다. 가짜 이미지의 경우, 모델의 성능을 높이기 위해 PSI(phase shift interferometry) 값이 각각 0.5, 0.7, 1.0인 이미지 5만 장씩을 조합하여 15만 장의 데이터셋을 구성하였다.

가짜 얼굴 이미지를 탐지하기 위해서는 진짜 및 가짜 얼굴 이미지의 특징을 추출하는 것이 필요하다. 본 논문은 대표적인 이미지 딥러닝 모델인 CNN을 사용하여 Fig. 1과 같은 가짜 얼굴 이미지 탐지 모델을 설계하였다. 제안된 모델은 정확도에 영향을 미칠 수 있는 다음과 같은 5개 개념들을 고려하여 설계되며, 이들의 모델 파라미터로서 변경 가능한 값으로 설정가능하다.

- *데이터 전처리(pre-processing)*

얼굴 이미지 데이터셋은 모두 컬러 이미지로서 RGB 세 개의 채널을 가지고 있다. 컴퓨터는 이미지를 인식할 때 비슷한 밝기를 가진 픽셀을 덩어리로 인식하므로, 이미지의 밝기, 명암 등에 따라 인식률에 차이가 발생한다. 본 연구에서는 이미지를 RGB로 사용하는 경우와 그레이스케일로 전처리하여 사용하는 경우를 나누어 학습 데이터를 구성하였다. 그레이스케일은 다중 채널인 컬러 이미지를 단일 채널인 흑백 이미지로 변환하여 데이터의 폭을 줄이는 역할을 한다.

- *컨볼루션 계층 수*

컨볼루션 계층(Convolution Layer)이 추가될수록 신경망은 학습을 통해 더욱 복잡한 형상을 인식하고, 이미지를 대표할 수 있는 공통적인 특징을 얻을 수 있게 된다. 본 연구에서는 컨볼루션 계층 수를 다르게 하여 정확도를 비교할 수 있도록 모델을 설계한다. 계층 수와 정확도가 비례하는 것은 아니기 때문에 계층 수를 다르게 하여 모델을 학습하면서 높은 성능을 보이는 계층 수를 찾는 것이 필요하다. 또한, 본 연구에서는 각 계층마다 맥스풀링(max pooling) 계층을 배치하여 데이터의 크기를 줄이고, 마지막에는 완전연결 계층(Fully-connected layer)을 통과시켜 최종 이미지를 분류한다.

- *커널(kernel) 크기*

커널 크기는 컨볼루션 계층에서 은닉층을 구성하기 위해 사용하는 가중치 개수를 의미한다. 커널의 크기가 홀수인 경우, 이전 모든 계층의 픽셀이 출력 픽셀 주위에 대칭으로 나타나기 때문에 짝수 크기의 커널보다 더 좋은 성능을 가진다. 또한, 큰 크기의 커널보다 작은 크기의 커널을 중첩해서 사용하면 더 작고 복잡한 특징을 추출할 수 있다. 일반적으로 3×3 크기의 커널을 주로 사용하지만 본 논문에서는 커널 크기에 따른 모델의 성능을 비교하기 위해 커널 크기를 변경할 수 있도록 한다.

- *커널 개수*

커널 개수는 해당 계층에 대한 입력 깊이를 결정

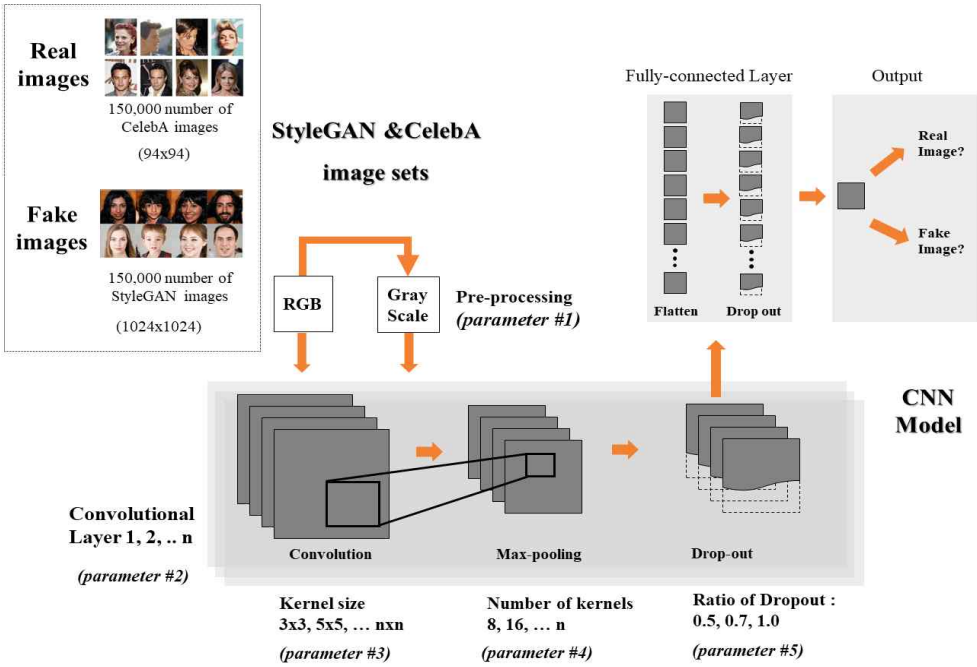


Fig. 1. Design of CNN-based StyleGAN image detection model considering 5 types of model parameters.

하는 요소로서 커널 개수만큼 뉴런 개수가 생성된다. 커널 개수는 모델의 균형을 맞추기 위해 각 계층에서 연산 시간 및 연산량을 고려하여 결정해야 한다. 커널 개수가 너무 많으면 시·공간 자원에서 모두 낭비를 초래하고, 너무 적으면 특징 학습량이 부족해진다. 본 논문에서는 커널 개수에 차이를 두어 커널 수에 따른 모델의 성능을 비교할 수 있게 한다.

- *드롭아웃 (Drop-out)*

드롭아웃은 입력된 데이터에 의존하여 학습할 때 발생할 수 있는 과적합을 방지하기 위한 방법으로서 하나의 컨볼루션 계층을 통과할 때마다 학습된 전체 뉴런 중에서 일부만 사용하고 일부는 버리는 방법이다. 드롭아웃 비율이 1에 가까울수록 버리는 뉴런이 적어지게 되고, 그만큼 학습 데이터에 대한 의존도가 높아져서 실제 평가(test) 데이터를 판단할 때 정확성이 떨어질 수 있다. 본 논문에서는 학습된 뉴런을 컨볼루션 계층 및 완전연결 계층에서 그대로 사용하는 경우, 컨볼루션 계층에서는 그대로 사용하고 완전연결 계층에서는 일부만 사용하는 경우, 그리고 컨볼루션 계층과 완전연결 계층에서 모두 일부만 사용하는 경우를 고려하여 모델을 설계한다.

4. 실험 결과 및 분석

4.1. 실험 시나리오

본 장에서는 3장에서 설명한 모델 파라미터에 변화를 주어 다양한 실험을 수행하고, 실험 결과 분석을 통해 모델 파라미터와 성능의 관계를 분석하고자 한다. 실험을 위한 학습 및 평가 데이터셋으로서 진짜 및 가짜 얼굴 이미지 데이터셋 총 30만 장 중, 70%는 학습 데이터셋(training set)으로 사용하고, 30%는 평가 데이터셋(testing set)으로 사용한다. Table 1은 제안된 모델의 파라미터 값을 고려하여 수행한 총 48개의 실험 시나리오 목록을 보여준다.

4.2 모델의 학습 및 실험 결과 분석

제안된 모델의 성능을 평가하기 위한 지표로서 TP(true positives), FN(false negatives), FP(false positives), TN(true negatives) 및 f1-score를 측정하였다. TP는 실제 참인 데이터를 참이라고 예측한 개수, FN은 실제 참인 데이터를 거짓으로 잘못 예측한 개수, FP는 실제 거짓인 데이터를 참으로 잘못 예측한 개수, TN은 실제 거짓인 데이터를 거짓으로 예측한 개수를 의미한다. 탐지한 모델의 성능을 판단

Table 1. 48 number of scenarios with the combination of model parameters

Scen-ario	Data prepro-cessing	Num. of Conv. layer	kernel size	Num. of kernels	Dropout			Scen-ario	Data prepro-cessing	Num. of Conv. layer	kernel size	Num of kernels	Dropout		
					Conv. layer		Fully-connected layer						Conv. layer		Fully-connected layer
					1	2							1	2	
1	Grayscale	1	3×3	8	0.7	-	0.5	25	Grayscale	1	5×5	8	0.7	-	0.5
2					1.0	-	0.7	26					1.0	-	0.7
3					1.0	-	1.0	27					1.0	-	1.0
4	RGB	1	3×3	8	0.7	-	0.5	28	RGB	1	5×5	8	0.7	-	0.5
5					1.0	-	0.7	29					1.0	-	0.7
6					1.0	-	1.0	30					1.0	-	1.0
7	Grayscale	1	3×3	16	0.7	-	0.5	31	Grayscale	1	5×5	16	0.7	-	0.5
8					1.0	-	0.7	32					1.0	-	0.7
9					1.0	-	1.0	33					1.0	-	1.0
10	RGB	1	3×3	16	0.7	-	0.5	34	RGB	1	5×5	16	0.7	-	0.5
11					1.0	-	0.7	35					1.0	-	0.7
12					1.0	-	1.0	36					1.0	-	1.0
13	Grayscale	2	3×3	8	0.7	0.7	0.5	37	Grayscale	2	5×5	8	0.7	0.7	0.5
14					1.0	1.0	0.7	38					1.0	1.0	0.7
15					1.0	1.0	1.0	39					1.0	1.0	1.0
16	RGB	2	3×3	8	0.7	0.7	0.5	40	RGB	2	5×5	8	0.7	0.7	0.5
17					1.0	1.0	0.7	41					1.0	1.0	0.7
18					1.0	1.0	1.0	42					1.0	1.0	1.0
19	Grayscale	2	3×3	16	0.7	0.7	0.5	43	Grayscale	2	5×5	16	0.7	0.7	0.5
20					1.0	1.0	0.7	44					1.0	1.0	0.7
21					1.0	1.0	1.0	45					1.0	1.0	1.0
22	RGB	2	3×3	16	0.7	0.7	0.5	46	RGB	2	5×5	16	0.7	0.7	0.5
23					1.0	1.0	0.7	47					1.0	1.0	0.7
24					1.0	1.0	1.0	48					1.0	1.0	1.0

하기 위해서는 정밀도(precision)와 검출율(recall)을 동시에 고려해야 하는데 f-socre가 바로 두 값을 하나의 값으로 표현한 지표이다. F-score 계산 시 정밀도에 가중치 베타값 1을 부여한 값이 바로 f1-score이며 f1-score는 아래 수식으로 표현된다.

$$f1-score = \frac{2\times precision \times recall}{precision + recall} \quad (1)$$

단, $precision = \frac{TP}{TP+FP}$ 및 $recall = \frac{TP}{FN+TP}$

실험은 CPU i7-9700 3.6GHz, GPU RTX 2070, 윈도우 10 운영체제에서 수행되었고, 모델은 텐서플로우 및 케라스를 활용하여 구현하였다. 모델의 활성함수는 'relu'가 사용되었고, 각 시나리오별로 약 40분의 학습시간이 소요되었다. Table 2는 실험을 통해 측정한 시나리오별 실험지표 결과를 보여주고 있으며 실험 결과, 48개 시나리오 모두 98% 이상의 탐지율을 보인 것을 확인할 수 있다.

48개의 시나리오는 제안된 모델의 파라미터 조합으로 개발되었기 때문에 시나리오별 실험 결과를 비교하면 모델 파라미터와 성능의 상관관계를 분석할 수 있다. 따라서 FN과 FP 분석을 통해 어떤 모델 파라미터들이 모델의 성능에 많은 영향을 미치는지 분석해보고자 한다. Fig. 2는 48개 시나리오를 FN 및 FP를 기준으로 분석한 그래프이다. 그래프는 FN 및 FP 각각 상위 12개(75th percentile), 하위 12개(25th percentile)에 해당하는 시나리오들에 대해, 설

Table 2. Experimental results by scenarios

Scenario	Real face images (45,000 images)		Fake face images (45,000 images)		F1-score	Scenario	Real face images (45,000 images)		Fake face images (45,000 images)		F1-score
	TP	FN	FP	TN			TP	FN	FP	TN	
1	44350	650	368	44632	0.99	25	6958	2042	1213	7787	0.82
2	44527	473	563	44437	0.99	26	7562	1438	1424	7576	0.84
3	44831	619	479	44521	0.99	27	7696	1304	1536	7464	0.84
4	44725	275	262	44738	0.99	28	44795	205	331	44669	0.99
5	44846	154	527	44473	0.99	29	44804	196	440	44560	0.99
6	44747	253	346	44654	0.99	30	44863	137	483	44517	0.99
7	44756	244	688	44312	0.99	31	7136	1864	804	8196	0.85
8	44829	171	1052	43948	0.99	32	7170	1830	959	8041	0.85
9	44673	327	364	44636	0.99	33	6629	2371	874	8126	0.82
10	44815	185	428	44572	0.99	34	44865	135	352	44648	0.99
11	44719	281	317	44683	0.99	35	44870	130	470	44530	0.99
12	44770	230	801	44199	0.99	36	44851	149	499	44501	0.99
13	44464	536	475	44525	0.99	37	44458	542	811	44189	0.98
14	44309	691	464	44536	0.99	38	44336	664	641	44359	0.99
15	44648	352	1036	43964	0.98	39	44095	905	535	44465	0.98
16	44747	253	552	44448	0.99	40	44729	271	521	44479	0.99
17	44809	191	513	44487	0.99	41	44681	319	478	44522	0.99
18	44885	115	1083	43917	0.99	42	44654	346	471	44529	0.99
19	44369	631	237	44763	0.99	43	44598	402	581	44419	0.99
20	44501	499	287	44713	0.99	44	44714	286	817	44183	0.99
21	44556	444	577	44423	0.99	45	44749	251	819	44181	0.99
22	44855	145	693	44307	0.99	46	44826	174	513	44487	0.99
23	44828	172	382	44618	0.99	47	44835	165	472	44528	0.99
24	44477	523	132	44868	0.99	48	44889	111	761	44239	0.99

정된 모델 파라미터 값을 유형별로 카운트하고 있다.

- *데이터 전처리*

FN은 진짜 얼굴 이미지를 가짜 이미지로 잘못 분류하는 상황이고, FP는 가짜 얼굴 이미지를 진짜 이미지로 잘못 분류하는 상황이다. 즉, FN 또는 FP가 높을수록 모델의 성능이 나쁘다고 할 수 있다. Fig. 2(a)에서 FN의 경우, 상위 12개에 속하는 시나리오 중, 그레이스케일 모델이 11개, 하위 12개 속하는 시나리오 중 RGB 모델이 11개로 큰 차이가 나는 것을 볼 수 있다. 이는 진짜 얼굴 이미지 탐지 시에는 RGB 이미지를 사용해야 탐지 정확성을 높일 수 있음을 보여준다. FP의 경우에도 상위 12개 시나리오에 그레이 스케일이 8개 포함되고, 하위 12개 시나리오에서는 RGB 모델이 7개 포함되는 것을 볼 수 있다. 이는 가짜 이미지 탐지 시에도 그레이스케일보다 RGB 모델 성능이 더 높은 것을 보여주는 결과이지만, FN에 비해 차이 값이 작은 것으로 보아 데이터 전처리 파라미터에 대한 민감도가 FN보다는 낮다고 할 수 있다.

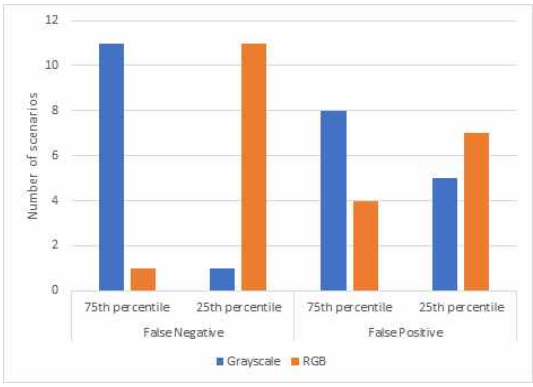
- 커널 개수

Fig. 2(d)의 FN의 경우, FN이 높은 상위 12개 시나리오 중, 커널 개수가 8개인 시나리오가 9개나 포함되는 것을 볼 수 있다. 반대로 하위 12개 시나리오 중에는 커널 개수가 16개인 시나리오가 9개 분포한다. 이는 진짜 이미지 탐지 시에는 커널 개수가 높을

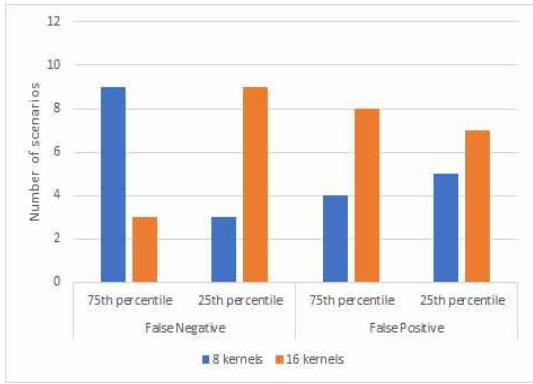
수록 진짜 이미지 탐지율이 향상된다는 것을 의미한다. 그러나 FP의 경우에는 상·하위 순위 모두에서 커널 개수가 16개인 시나리오가 더 많은 것으로 보아 가짜 이미지 탐지 시에는 커널의 개수가 탐지 성능에 큰 영향을 미치지 않는 것으로 판단할 수 있다.

- *컨볼루션 계층 수 및 커널 크기*

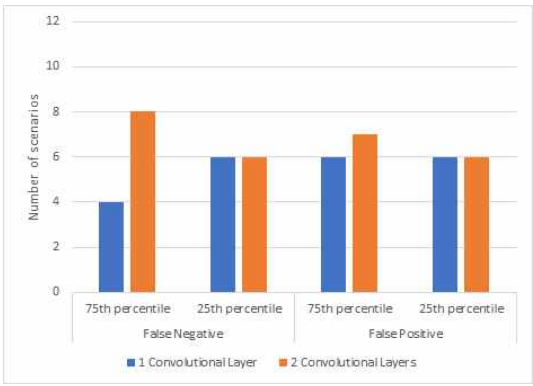
Fig. 2의 (c)와 (d)는 컨볼루션 계층 수 및 커널 크기에 따른 FN 및 FP 값을 보여준다. 모델 파라미터가 성능에 영향을 미친다면 그래프의 계열별 수치 차이가 크게 발생하거나 상·하위 결과에서 서로 상반된



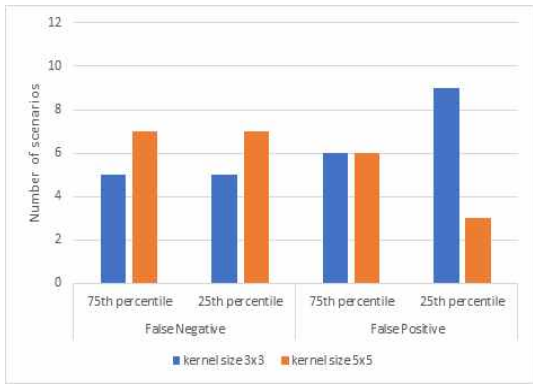
(a) pre-processing



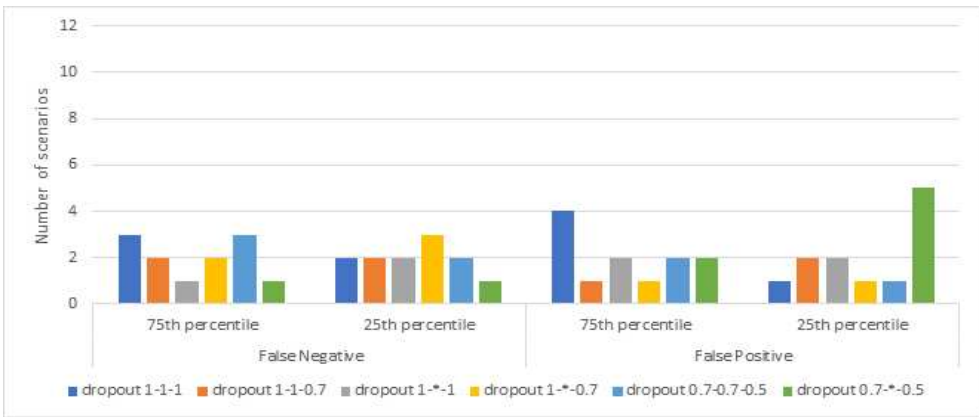
(b) number of kernels



(c) number of convolutional layers



(d) kernel size



(e) ratio of dropout

Fig. 2. Analysis of false negatives and false positives by model parameters.

결과 패턴이 보여져야 한다. 그러나 계열별 그래프가 완만할 뿐 아니라, 상·하위 결과에서 서로 상반된 결과 패턴도 나타나지 않는 것으로 보아, 컨볼루션 계층 수 및 커널 크기는 다른 모델 파라미터에 비해 성능에 미치는 영향이 미비하다고 할 수 있다.

- 드롭아웃

Fig. 2(e)에서 FN의 경우에는 드롭아웃 변화에 따른 계열별 차이가 두드러지지 않고 완만한 것으로 보아 진짜 이미지 탐지 시에는 드롭아웃이 모델의 성능에 큰 영향을 미치지 않는다고 판단할 수 있다. 그러나 FP 상위 12개의 경우, 드롭아웃 비율이 1(컨볼루션계층 1)-1(컨볼루션계층 2)-1(완전연결계층)인 시나리오의 분포가 가장 많은 것을 볼 수 있는데 이는 학습된 뉴런을 일부 제거하지 않고 모두 그대로 학습에 계속 반영하여 과적합이 발생하였음을 알 수 있다. 반대로 FP의 하위 12개 시나리오에서는 드롭아웃 비율이 0.7(컨볼루션계층 1)-0.5(완전연결계층)로 설정된 시나리오들이 가장 많이 분포하는 것으로 보아 드롭아웃을 통해 과적합을 방지함으로써 가짜 이미지 탐지율을 높일 수 있음을 확인하였다.

5. 결론 및 향후 연구계획

본 논문은 StyleGAN에 의해 생성된 가짜 얼굴 이미지를 탐지하기 위한 CNN 기반의 딥러닝 모델을 제안하였다. 제안된 모델은 학습할 이미지를 RGB에서 그레이스케일로 변환하는 데이터 전처리 과정, 컨볼루션 계층 모델링에 필요한 컨볼루션 계층 수, 커널 크기, 커널 개수, 그리고 드롭아웃을 모델 파라미터로 가지도록 설계하였다. 또한, 모델 파라미터들의 여러 조합을 고려하여 총 48개의 모델 시나리오를 개발하고, 진짜 얼굴 이미지 15만장 및 StyleGAN으로 생성된 가짜 얼굴 이미지 15만장에 대해 학습 및 평가를 수행하여 각 시나리오별 탐지 성능을 측정하였다. 48개 시나리오의 f1-score는 모두 0.98이상으로서 제안된 모델이 StyleGAN 얼굴 이미지 대부분을 정확하게 탐지하고 있음을 확인하였다. 또한, 이미지를 잘못 예측한 FN 및 FP를 기준으로 상·하위 25%에 속하는 시나리오들을 선별하여 모델 파라미터를 분석한 결과, 이미지를 그레이스케일로 전처리하지 않았을 때, 즉, RGB 이미지를 사용하였을 때 진짜 및 가짜 얼굴 이미지 탐지율이 모두 향상되었음을 확인하였다. 또한, 커널 개수가 많을수록 진짜 얼굴 이미지 탐지에 좋은 성능을 보이는 것을 확인할 수 있었다. 그러나 컨볼루션 계층 수 및 커널 사이즈는 모델의 성능에 큰 영향을 미치지 못하고, 드롭아웃을 통해 학습된 뉴런의 일부를 제거하면 가짜 얼굴 이미지 탐지 시 성능을 향상 시킬 수 있음을 확인하였다. 본 논문에서는 실험결과에 대한 고찰로서 진짜 얼굴이미지의 해상도에 따라 실험결과가 어떻게 달라지는지 추가로 분석해 보았다. 본문에서 94x94인 CelebA 진짜 이미지 15만 장과 StyleGAN 가짜 이미지 15만 장을 사용했다면, 추가실험에서는 해상도가 1024×1024인 CelebA-HQ 이미지 3만장과 StyleGAN 이미지 3만장을 활용하여 48개의 동일한 시나리오로 모델을 학습하였다. 본문에서 제시한 실험과 비교해 본 결과, 모델 파라미터 중, 이미지 전처리, 커널 사이즈, 컨볼루션 계층 수, 드롭아웃율에 있어서는 해상도가 낮았을 때와 모두 같은 결과를 얻은 것을 확인할 수 있었다. 단, 커널 개수와 관련해서는 기존에 발견하지 못했던 추가적인 특성을 확인할 수 있었다. 즉, 해상도가 낮았을 때에는 커널의 개수가 많을수록 진짜 이미지 탐지 성능이 더 높아졌지만, 가짜 이미지 탐지 시에는 커널의 개수가 성능에 미치는 영향을 발견하지 못했다. 그러나 진짜 이미지의 해상도가 높은 추가실험에서는 진짜 이미지 뿐만 아니라, 가짜 이미지 탐지 시에도 커널의 개수가 많을수록 성능이 향상되는 것을 확인할 수 있었다. 따라서 향후 연구에서는 학습 이미지 데이터셋의 속성에 따른 모델 파라미터와 성능의 관계를 심층 분석할 예정이며, StyleGAN 이미지 뿐 아니라, 다른 종류의 GAN으로 생성한 가짜 이미지 탐지에도 높은 정확성을 갖도록 제시된 모델을 확장할 예정이다. 본 논문에서 개발한 모델은 딥러닝 기술을 이용한 딥페이크 이미지 탐지 기술로 활용되어 불법 콘텐츠 제작, 이미지 위조를 통한 명예훼손 등 범죄 예방에도 기여할 수 있을 것으로 기대한다.

REFERENCE

[1] A. Radford, L. Metz, and C. Soumith, “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks,” *Computer Science, Machine Learning*, *arXiv Preprint arXiv:1511.06434*, 2015.

[2] D. Berthelot, T. Schumm, and L. Metz, "Began: Boundary Equilibrium Generative Adversarial Networks," *Computer Science, arXiv Preprint arXiv:1703.10717*, 2017.

[3] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al., "Generative Adversarial Nets," *Advances in Neural Information Processing Systems,* pp. 2672-2680, 2014.

[4] T. Karras, S. Laine, and T. Aila, "A Style-based Generator Architecture for Generative Adversarial Networks," *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4401-4410, 2019.

[5] T. Karras, T. Aila, S. Laine, and J. Lehtinen, ARRAS, "Progressive Growing of Gans for Improved Quality, Stability, and Variation," *Computer Science, Neural and Evolutionary Computing*, *arXiv Preprint arXiv:1710.10196*, 2017.

[6] D.H. Kim, C.S. Wan, and K.S. Yeong, "Deep Learning Based Fake Face Detection," *Journal of the Korea Industrial Information Systems Research*, Vol. 23, No. 5, pp. 9-17, 2018.

[7] C.C. Hsu, Y.X. Zhuang, and C.Y. LEE, "Deep Fake Image Detection Based on Pairwise Learning," Preprints 2019, 201905.0013.v1, 2019.

[8] L.M. Dang, S.I. Hassan, S.Y. Im, J.C. Lee, S.J. Lee, and H.J. Moon, et al., "Deep Learning Based Computer Generated Face Identification Using Convolutional Neural Network," *Applied Sciences*, Vol. 8, No. 2610, doi:10.3390/app8122610, 2018.

[9] C.C. Hsu, C.Y. LEE, and Y.X. Zhuang, "Learning to Detect Fake Face Images in the Wild," *Proceeding of IEEE International Symposium on Computer, Consumer and Control Conference*, pp. 388-391, 2018.

[10] H. Li, B. Li, S. Tan, and J. Huang, "Detection of Deep Network Generated Images Using Disparities in Color Components," *Computer Science*, *arXiv Preprint arXiv:1808.07276*, 2018.

[11] T.D. Nhu, I.S. Na, and S.H. KIM, "Forensics Face Detection From GANs Using Convolutional Neural Network," Proceeding of 2018 International Symposium on Information Technology Convergence(ISITC 2018), 2018.

[12] S.K. Yarlagadda, D. Guera, P. Bestagini, F.M. Zhu, S. Tubaro, and E.J. Delp, "Satellite Image Forgery Detection and Localization Using GAN and One-class Classifier," *Electronic Imaging*, Vol. 7, pp. 1-9, 2018.

[13] J.Y. Zhu, T. Park, P. Isola, and A.A. Efros, "Unpaired Image-to-image Translation Using Cycle-consistent Adversarial Networks," *Proceeding of the IEEE International Conference on Computer Vision*, pp. 2223-2232, 2017.

[14] Y.J. Choi, M.J. Choi, M.Y. Kim, J.W. Ha, S.H. Kim, and J.G. Choo, "Stargan: Unified Generative Adversarial Networks for Multi-domain Image-to-image Translation," *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8789-8797, 2018.

[15] L. Nataraj, T.M. Mohammed, S. Chandrasekaran, A. Flenner, J.H. Bappy, A.K. Roy-Chowdhury, et al., "Detecting GAN Generated Fake Images Using Co-occurrence Matrices," *Computer Science, Computer Vision and Pattern Recognition*, *arXiv Preprint arXiv:1903.06836*, 2019.

[16] X. Xuan, B. Peng, W. Wang, and J. Dong, "On the Generalization of GAN Image Forensics," *Proceeding of Chinese Conference on Biometric Recognition, Springer, Cham,* pp. 134-141, 2019.

[17] S. Tariq, S.Y. Lee, H.Y. Kim, Y.J. Shin, and S.S. Woo, "GAN Is a Friend or Foe?: a Framework to Detect Various Fake Face Images," *Proceeding of the 34th ACM/SIGAPP Symposium on Applied Computing,* pp. 1296-1303, 2019.

[18] Yang Yang, Lei Zheng, Jiawei Zhang, Qingcai Cui, Zhoujun Li and Philip S. Yu, "TI-CNN: Convolutional Neural Networks for Fake News Detection," *arXiv Preprint arXiv:1806.00749*, 2018.

[19] X. Zhang, S. Karaman, and S.F. Chang, "Detecting and Simulating Artifacts in GAN Fake Images," *Computer Science, Computer Vision and Pattern Recognition, arXiv Preprint arXiv:1907.06515*, 2019.

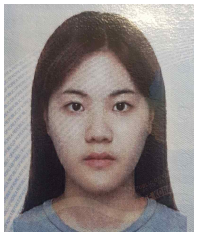
[20] S.Y. Rhyou, H.J. Kim, and K.A. Cha, "Development of Access Management System Based on Face Recognition Using ResNet," *Journal of Korea Multimedia Society*, Vol. 22, Issue 8, pp. 823-831, 2019.

김 지 연

2007년 2월 서울여자대학교 정보보호공학과(공학사)
2013년 8월 서울여자대학교 컴퓨터학과(이학박사)
2014년 3월~2017년 8월 Carnegie Mellon University 박사후연구원
2019년 3월~현재 서울여자대학교 소프트웨어교육혁신센터 전담교수
관심분야 : 네트워크 보안, 인공지능, 클라우드 보안, 사물인터넷보안

홍 승 아

2018년 3월~현재 서울여자대학교 정보보호학과
관심분야 : 정보보호, 인공지능, 빅데이터



김 하 민

2018년 3월~현재 서울여자대학교 정보보호학과
관심분야 : 정보보호, 인공지능, 빅데이터