Journal of Korea Multimedia Society Vol. 24, No. 12, December 2021(pp. 1624-1631)
https://doi.org/10.9717/kmms.2021.24.12.1624

일반화 능력이 향상된 CNN 기반 위조 영상 식별

이정한†, 박한훈††

CNN-Based Fake Image Identification with Improved Generalization

Jeonghan Lee†, Hanhoon Park††

ABSTRACT

With the continued development of image processing technology, we live in a time when it is difficult to visually discriminate processed (or tampered) images from real images. However, as the risk of fake images being misused for crime increases, the importance of image forensic science for identifying fake images is emerging. Currently, various deep learning-based identifiers have been studied, but there are still many problems to be used in real situations. Due to the inherent characteristics of deep learning that strongly relies on given training data, it is very vulnerable to evaluating data that has never been viewed. Therefore, we try to find a way to improve generalization ability of deep learning-based fake image identifiers. First, images with various contents were added to the training dataset to resolve the over-fitting problem that the identifier can only classify real and fake images with specific contents but fails for those with other contents. Next, color spaces other than RGB were exploited. That is, fake image identification was attempted on color spaces not considered when creating fake images, such as HSV and YCbCr. Finally, dropout, which is commonly used for generalization of neural networks, was used. Through experimental results, it has been confirmed that the color space conversion to HSV is the best solution and its combination with the approach of increasing the training dataset significantly can greatly improve the accuracy and generalization ability of deep learning-based identifiers in identifying fake images that have never been seen before.

Key words: Fake Image Identification, Generative Adversarial Networks, Image Forensics, CNN, Generalization

1. 서 론

컴퓨터 학습을 통해 실존하지 않지만 존재할 법한 영상을 제작하는 연구가 활발히 진행되고 있다. 딥러닝(deep learning) 기술의 발전과 함께 더 좋은 성능의 영상 생성 기법[1]들이 개발되고 있다. 예를 들어, 해상도가 낮은 영상의 해상도를 학습을 통해 높이거나[2] 기존 영상의 화풍, 계절, 사물까지 학습을 통해 바꿔버리는 기법[3]도 존재한다. 이처럼, 학습 기반의 영상 생성 기술은 다양한 방식으로 세분화되어 그 목적에 걸맞는 우수한 성능을 자랑하고 있다.

영상 생성 기술의 발전에 따라 실제(real) 영상을 모방한 영상의 실존 여부는 육안으로 구분하기엔 더 모호하고, 대부분의 기술들이 실제 영상에 부분적인 변화만을 주기 때문에 컴퓨터를 사용해서 구분하는 것도 쉽지 않다. 그러나, 생성된 영상을 범죄에 악용

※ Corresponding Author : Hanhoon Park, Address: (48513) Yongso-ro 45, Nam-gu, Busan, Korea, TEL : +82-51-629-6225, FAX : +82-51-629-6210, E-mail : hanhoon.park@pknu.ac.kr
Receipt date : Nov. 10, 2021, Revision date : Dec. 9, 2021
Approval date : Dec. 21, 2021

† Dept. of Electronic Engineering, Pukyong National University (E-mail : jeonghan_lee@pukyong.ac.kr)
†† Dept. of Electronic Engineering, Pukyong National University
※ This work was supported by a Research Grant of Pukyong National University (2021).

되는 사례가 보도되고 있어, 실제 영상과 생성된 영상(즉, 위조(fake) 영상)을 구분하기 위해 신경망 기반의 컴퓨터 학습을 이용하는 연구[4,5]도 함께 진행되고 있다.

그러나, 위조 영상을 식별하기 위한 기존 방법들은 학습 시 사용된 영상과 유사한 콘텐츠를 가지거나 같은 생성 네트워크를 기반으로 제작된 영상에 대해서만 식별이 가능하다는 제약을 가진다. 그러나, SNS와 인터넷 환경 같은 실제 상황에선 학습 시 사용했던 영상과 생성 네트워크 모델이나 콘텐츠가 전혀 다른 영상을 식별해야 하는 경우가 흔하다. 그러므로, 본 논문에서는 이러한 상황을 고려하여 일반화(generalization) 능력이 향상된 위조 영상 식별 방법을 제안한다. 이를 위해 위조 영상 식별 성능이 우수한 딥러닝 모델을 선정하고, 일반화 능력 향상을 위한 방안들을 적용한 후 각 방안의 성능을 실험적으로 검증함으로써 일반화 능력이 가장 우수한 조합을 찾는다.

본 논문의 구성은 다음과 같다. 2 장에서는 영상 생성을 위한 GAN(Generative Adversarial Network)과 딥러닝 기반의 영상 분류에 대해서 간략히 소개하고, 3 장에서는 일반화 능력을 향상시키기 위한 방법에 대해 설명한다. 4 장에서는 위조 영상 식별을 위해 사용된 딥러닝 모델을 소개하고, 실험을 통해 일반화 능력 향상 방법을 적용했을 때의 성능 변화를 분석한다. 5 장에서는 결론과 향후 연구 방향을 제시한다.

2. 관련 연구

2.1 GAN

GAN[6]은 영상을 생성하는 신경망의 일종으로, 입력 데이터를 학습하여 최대한 유사한 표본을 생산하는 것이 목적이다. GAN의 생성기(generator) 신경망을 통해 생성된 데이터는 분류기(discriminator) 신경망을 통해 진짜인지 가짜인지 분류하고 생성기가 이 분류정보를 참고하면서 성능을 향상시키는 것이 적대적인 생성 네트워크, GAN의 핵심이다. GAN으로 생성된 영상은 생성기의 내부 전결합층을 통과해 최종적으로 RGB 특성값만을 고려하여 구성된 디지털 영상이 된다. 최근 네트워크 구조 및 사용 목적에 따라 다양한 형태의 GAN 모델이 제안되고 있으나, 본 논문에서는 가장 보편적으로 사용되면서 이미 다양한 콘텐츠를 이용하여 많은 양의 생성 영상을 제공하는 ProGAN[7]을 위조 영상의 생성 모델로 사용한다.

2.2 딥러닝 기반의 영상 분류

신경망의 일종인 합성곱 신경망(Convolutional Neural Network, CNN)은 영상의 시각적 분석에 주로 적용된다. 최근 CNN은 네트워크 모델을 최대한 깊게(deep) 구성하여, 더욱 어려운 과제까지 수행 가능하도록 발전하였다. 그러나, 단순히 CNN을 깊게 발전시키면서 특정 정보에 편향적으로 학습하게 되고, 처음 접하는 정보에 대하여 취약하다는 단점도 공존했다.

CNN에 기반한 딥러닝 기술은 영상 분류에서 탁월한 성능을 보여주고 있으며, 다양한 분야에서 차용해 폭넓게 사용되고 있다. 예를 들어, 실제 카메라로 촬영한 영상이 카메라의 기종에 따라 고유한 흔적을 남기듯이 영상 생성 모델의 내재적인 흔적을 추적하여 구분[8,9]하거나, 육안으로는 보기 힘든 고주파 인공물을 찾기 위해 영상 법의학에서는 이미 널리 사용되는 고주파 통과 필터를 적용[10]하는 연구도 함께 진행되었다. 또한, 영상 추적에 흔히 적용되는 퓨샷 러닝(few-shot learning)을 적용해 매우 적은 증거만으로도 가짜 영상을 분류해 내는 방법[11,12] 등도 제안되었다. 그러나 기존 딥러닝 기반의 영상 분류는 단순히 동일한 조건 하에 있는 데이터를 사용하여 학습되고 평가되거나, 학습 시 사용된 데이터에 과적합(over-fitting)되는 문제가 있다. 그러므로, 기존 딥러닝 기반 영상 분류 방법들의 일반화 능력을 향상시키기 위한 기술이 요구된다.

3. 일반화 능력 향상

본 논문은 GAN에 의해 생성된 위조 영상을 식별하는 것을 목적으로 하며, 딥러닝 기반 위조 영상 분류 방법의 일반화 능력을 개선하기 위해 기존의 다양한 연구 결과를 바탕으로 일반화 능력 향상에 효과적인 방법들의 성능을 실험적으로 검증하고, 최적의 조합을 찾는다.

먼저, 딥러닝 모델이 영상 콘텐츠에 기반하여 실제와 위조 영상을 분류하지 않도록 다양한 콘텐츠를

혼합하여 학습시킴으로써 일반화를 유도하는 방법을 사용한다. 즉, 학습에 사용되는 영상의 카테고리 수를 점진적으로 늘리면서 학습했을 때, 학습 때와 다른 카테고리 영상을 분류할 때 일반화 능력이 향상되는지 실험을 통해 검증한다.

다음으로, GAN에서 생성된 영상이 단순히 RGB 채널 기반으로 생성된다는 점을 이용하여 인위적으로 생성된 위조 영상과 현실에서 습득된 실제 영상이 HSV와 YCbCr 색 공간에선 RGB 채널에서 나타나지 않는 왜곡이 존재하며, 이는 딥러닝 기반 위조 영상 식별 방법의 일반화 능력을 개선할 수 있다는 이전 연구[13] 결과를 참고해 단순히 디지털 영상 표현에 사용되는 RGB뿐만 아니라 HSV와 YCbCr와 같은 색 공간을 사용하여 위조 영상을 식별하는 방법의 성능을 분석한다.

마지막으로, 학습 시 보지 않은 데이터에 대한 식별 정확도를 높이기 위해 CNN의 일반화 능력 향상을 위한 보편적인 방법인 드롭아웃(dropout)의 비율을 조정하면서 성능 변화를 분석한다.

결과적으로 가장 좋은 결과를 보이는 방법들을 이용해 위조 영상 식별을 위한 CNN 모델에 적용한 최종 모델을 구축하고, 뛰어난 성능으로 기존 연구에서 위조 영상 식별에 흔히 사용되는 Xception 모델[14]과 성능을 비교한다.

4. 실험 및 분석

4.1 실험 환경

4.1.1 데이터셋

실험에 사용된 데이터는 LSUN[15]에서 임의로 선택한 11개의 카테고리 영상(cat, church_outdoor, train, airplane, bus, cow, bridge, bedroom, classroom, restaurant, sheep)과 이를 이용하여 ProGAN으로 생성된 영상을 각각 “Real”과 “Fake”로 설정하여 각각 25,000장씩 구성하였다.

본 논문에서는 딥러닝 모델이 영상 콘텐츠에 기반하여 실제와 위조 영상을 분류하지 않도록 다양한 콘텐츠를 혼합하여 학습시킴으로써 일반화를 유도하는 방법을 고안하였다. 학습 데이터셋은 레이블 당 최대 25,000장이 넘지 않도록 같은 비율로 카테고리 수를 점진적으로 늘렸고, 평가 데이터셋은 각각 카테고리의 레이블당 2,000장씩 구성하였다. 이후, 학습에 포함된 카테고리 영상에 대한 분류 결과와 포함되지 않은 카테고리 영상에 대한 분류 결과를 따로 평균을 내어 그 성능을 확인하였다. “sheep” 카테고리의 경우 한번도 보지 않은 대표 평가 데이터셋으로 따로 분리해 두어 훈련 데이터셋에 포함되지 않게 했다.

4.1.2 딥러닝 모델

본 논문에서 사용된 딥러닝 모델은 전처리 단계에서 고주파 통과 필터(HPF)를 사용하는 Pelee[16]이다. Pelee는 실시간 객체 탐지를 위해 기존의 CNN 모델과 차별화를 둔 네크워크로 최대한 효율적으로 빠른 속도를 내기 위해 최적화된 구조를 지닌 특징을 가지고 있다. 그럼에도 불구하고 분류정확도도 높기 때문에 영상 스테그아날리시스[17]을 비롯한 디지털 포렌식 분야에서 사용되어왔다. 또한, 고주파 통과 필터를 사용한 이유는 영상에 포함된 작은 변화나 실제 영상과 위조 영상 사이의 미묘한 차이 등을 증폭시켜 분류나 식별 정확도를 높여주기 때문이다.

4.1.3 실험 방법

일반화 능력을 분석하기 위해 기본 딥러닝 모델에 다양한 조건을 변경하면서 실험을 진행하였다. 먼저, 카테고리 수를 늘리면서 다양한 콘텐츠를 포함하도록 구성된 학습 데이터셋이 일반화 능력 향상에 기여하는지 확인하기 위해 점진적으로 카테고리를 추가하면서 학습하고, 학습 시 사용하지 않은 카테고리 영상들에 대해서 평가하는 실험을 진행하였다. 두 번째로, 영상의 색 공간에 따른 일반화 능력의 차이를 확인하기 위해 입력 영상을 RGB, HSV, YCbCr 색 공간으로 표현했을 때의 위조 영상 식별 정확도를 비교하는 실험을 진행하였다. 또한, 드롭아웃을 제일 마지막 층에 적용하여 일반화 능력의 변화를 분석하였다. 드롭아웃 비율은 0.2와 0.5로 설정하였다. 마지막으로, 다양한 조건을 가진 실험에서 가장 우수했던 경우를 조합하여 모델을 구성하였고, 기존 위조 영상 식별 연구에서 보편적으로 사용되고 있는 Xception과 식별 정확도를 비교하는 실험을 진행하였다.

4.2 학습 시 사용된 영상 카테고리 수에 따른 위조 영상 식별 정확도

앞서 언급한 것과 같이 첫 번째 실험으로 학습 데

Table 1. Fake image identification rates of Pelee+HPF depending on the number of image categories used in the training step.

Train	Test											mean ACC (Trained)	mean ACC (Not Trained)
	cat	church	train	airplane	bus	cow	bridge	bedroom	classroom	restaurant	sheep		
Single object(cat)	0.9985	**0.5**	**0.9395**	**0.947**	**0.6352**	**0.9648**	**0.5**	**0.5005**	**0.5**	**0.5**	**0.9305**	0.9985	**0.6918**
Mix_2 (+ church)	0.9985	0.9985	**0.8478**	**0.964**	**0.5505**	**0.9545**	**0.9492**	**0.9698**	**0.9572**	**0.824**	**0.9188**	0.9985	**0.8818**
Mix_3 (+ train)	0.914	0.9612	0.9428	**0.738**	**0.7658**	**0.9602**	**0.8965**	**0.865**	**0.962**	**0.9042**	**0.9652**	0.9393	**0.8821**
Mix_4 (+ airplane)	0.9995	0.999	0.9982	0.996	**0.6862**	**0.9888**	**0.9685**	**0.9675**	**0.9775**	**0.9315**	**0.9888**	0.9982	**0.9283**
Mix_5 (+ bus)	0.9985	0.9995	0.9988	0.9958	0.998	**0.9952**	**0.9458**	**0.9525**	**0.965**	**0.867**	**0.993**	0.9981	**0.9531**
Mix_6 (+ cow)	0.9958	0.9938	0.9938	0.9845	0.9858	0.995	**0.8825**	**0.8932**	**0.8962**	**0.9078**	**0.9828**	0.9914	**0.9125**
Mix_7 (+ bridge)	0.9995	0.997	0.9992	0.9988	0.999	0.9988	0.9985	**0.993**	**0.99**	**0.9518**	**0.9605**	0.9987	**0.9738**
Mix_8 (+ bedroom)	0.9982	0.9975	0.999	0.9985	0.999	0.999	0.9985	0.9978	**0.996**	**0.929**	**0.9735**	0.9984	**0.9662**
Mix_9 (+ classroom)	0.994	0.9965	0.9942	0.9952	0.9922	0.9945	0.9992	0.998	0.9988	**0.9895**	**0.9225**	0.9958	**0.956**
Mix_10 (+ restaurant)	0.998	0.9892	0.9965	0.9958	0.9902	0.997	0.9828	0.994	0.9858	0.9788	**0.9328**	0.9908	**0.9328**

이터셋에 영상 카테고리를 추가하는 과정을 점진적으로 반복하며, 학습 데이터셋에 포함된 콘텐츠의 양에 따라 학습에 포함되지 않은 영상을 식별할 때의 정확도가 바뀌는 양상을 확인하였다.

Table 1은 학습 데이터셋의 양은 동일하게 유지하면서 10개의 영상 카테고리를 점진적으로 추가함으로써, 모든 컨텐츠의 비율이 동일하도록 Real/Fake의 이진 분류 학습 데이터셋을 구성하여 학습했을 때의 식별 정확도 결과를 보여준다. 학습한 네트워크를 호출해 각 카테고리 영상으로 이루어진 평가 데이터셋을 사용해 모두 평가하고 학습에 포함된 데이터와 포함되지 않은 데이터를 분리해 각각 평균을 계산하였다. 학습에 사용된 카테고리에 속하는 영상들에 대해서는 식별 정확도가 매우 높지만(표의 대각선 아래 값들), 학습에 사용되지 않은 카테고리에 속하는 영상들에 대해서는 식별 정확도가 크게 떨어지는 것(표의 대각선 위 값들)을 확인할 수 있다.

Fig. 1은 학습에 포함되지 않은 영상들에 대해 위조 영상 식별 정확도를 나타낸 것으로, 학습 데이터셋에 포함된 영상의 카테고리(또는 콘텐츠)가 더 많을 경우 학습에 포함되지 않은 영상들에 대한 평균적인 정확도가 증가하여 일반화 능력이 향상된다는 것을 추세선을 통해 알 수 있었다. 그러나, 카테고리(또는 콘텐츠)가 더 많음에도 정확도가 오히려 하락하는 구간도 존재하여 일관적인 일반화 능력의 향상을 확인하기는 어려웠다.

4.3 색 공간 차이에 따른 위조 영상 식별 정확도

입력 영상을 HSV, YCbCr 색 공간으로 변환한 후 4.2절과 동일한 실험을 진행하였다. Table 2와 3은 각각 HSV와 YCbCr 색 공간에서 점진적으로 영상 카테고리를 추가해 학습했을 때 식별 정확도를 보여준다.

Fig. 2는 색 공간을 다르게 하여 실험을 진행했을 때 식별 정확도의 변화 양상이 어떻게 달라지는지를 그래프로 나타낸 것이다. RGB 색 공간을 사용할 때보다 HSV나 YCbCr 색 공간을 사용했을 때 일반적

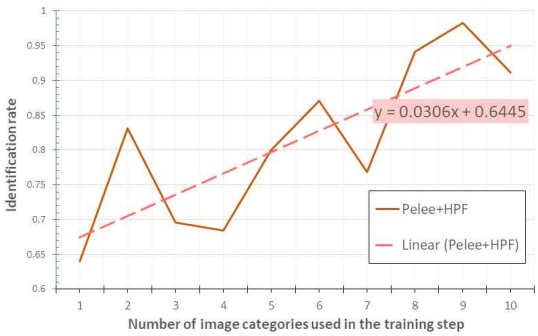

Fig. 1. Change in fake image identification rates for the images that are not included in the image categories used in the training step as the number of image categories used in the training step increases.

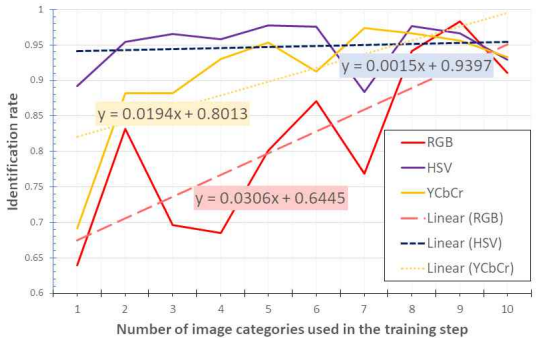

Fig. 2. Change in fake image identification rates for the images that are not included in the image categories used in the training step when using different color spaces.

Table 2. Fake image identification rates of Pelee+HPF when the image color space was changed to HSV.

Train	Test											mean ACC (Trained)	mean ACC (Not Trained)
	cat	church	train	airplane	bus	cow	bridge	bedroom	classroom	restaurant	sheep		
Single object(cat)	0.9898	**0.8822**	**0.9398**	**0.9722**	**0.8645**	**0.8332**	**0.912**	**0.9542**	**0.91**	**0.8295**	**0.8238**	0.9898	**0.8921**
Mix_2 (+ church)	0.9935	0.9975	**0.9458**	**0.9605**	**0.8878**	**0.9195**	**0.9808**	**0.9925**	**0.9955**	**0.9858**	**0.9218**	0.9955	**0.9544**
Mix_3 (+ train)	0.9945	0.9955	0.9952	**0.9918**	**0.9782**	**0.9078**	**0.9742**	**0.995**	**0.9988**	**0.9628**	**0.9182**	0.9951	**0.9658**
Mix_4 (+ airplane)	0.9582	0.9925	0.9898	0.9878	**0.9902**	**0.9465**	**0.9708**	**0.9848**	**0.9825**	**0.916**	**0.9142**	0.9821	**0.9578**
Mix_5 (+ bus)	0.9905	0.9918	0.9968	0.994	0.998	**0.9795**	**0.9755**	**0.9932**	**0.9925**	**0.953**	**0.9692**	0.9942	**0.9772**
Mix_6 (+ cow)	0.9958	0.9942	0.9975	0.9955	0.9968	0.9958	**0.979**	**0.9882**	**0.9985**	**0.9518**	**0.962**	0.9959	**0.9759**
Mix_7 (+ bridge)	0.9545	0.896	0.9795	0.8978	0.9288	0.9448	0.9378	**0.8248**	**0.9595**	**0.8232**	**0.9272**	0.9342	**0.8837**
Mix_8 (+ bedroom)	0.981	0.99	0.9942	0.9905	0.9865	0.983	0.9865	0.9875	**0.9972**	**0.962**	**0.97**	0.9874	**0.9764**
Mix_9 (+ classroom)	0.9695	0.9932	0.9888	0.9868	0.9945	0.973	0.9818	0.9952	0.9975	**0.9768**	**0.9562**	0.9867	**0.9665**
Mix_10 (+ restaurant)	0.986	0.9955	0.996	0.9968	0.9922	0.9762	0.994	0.998	0.9992	0.9965	**0.9292**	0.993	**0.9292**

Table 3. Fake image identification rates of Pelee+HPF when the image color space was changed to YCbCr.

Train	Test											mean ACC (Trained)	mean ACC (Not Trained)
	cat	church	train	airplane	bus	cow	bridge	bedroom	classroom	restaurant	sheep		
Single object(cat)	0.9985	**0.5**	**0.9395**	**0.947**	**0.6352**	**0.9648**	**0.5**	**0.5005**	**0.5**	**0.5**	**0.9305**	0.9985	**0.6918**
Mix_2 (+ church)	0.9985	0.9985	**0.8478**	**0.964**	**0.5505**	**0.9545**	**0.9492**	**0.9698**	**0.9572**	**0.824**	**0.9188**	0.9985	**0.8818**
Mix_3 (+ train)	0.914	0.9612	0.9428	**0.738**	**0.7658**	**0.9602**	**0.8965**	**0.865**	**0.962**	**0.9042**	**0.9652**	0.9393	**0.8821**
Mix_4 (+ airplane)	0.9995	0.999	0.9982	0.996	**0.6862**	**0.9888**	**0.9685**	**0.9675**	**0.9775**	**0.9315**	**0.9888**	0.9982	**0.9283**
Mix_5 (+ bus)	0.9985	0.9995	0.9988	0.9958	0.998	**0.9952**	**0.9458**	**0.9525**	**0.965**	**0.867**	**0.993**	0.9981	**0.9531**
Mix_6 (+ cow)	0.9958	0.9938	0.9938	0.9845	0.9858	0.995	**0.8825**	**0.8932**	**0.8962**	**0.9078**	**0.9828**	0.9914	**0.9125**
Mix_7 (+ bridge)	0.9995	0.997	0.9992	0.9988	0.999	0.9988	0.9985	**0.993**	**0.99**	**0.9518**	**0.9605**	0.9987	**0.9738**
Mix_8 (+ bedroom)	0.9982	0.9975	0.999	0.9985	0.999	0.999	0.9985	0.9978	**0.996**	**0.929**	**0.9735**	0.9984	**0.9662**
Mix_9 (+ classroom)	0.994	0.9965	0.9942	0.9952	0.9922	0.9945	0.9992	0.998	0.9988	**0.9895**	**0.9225**	0.9958	**0.956**
Mix_10 (+ restaurant)	0.998	0.9892	0.9965	0.9958	0.9902	0.997	0.9828	0.994	0.9858	0.9788	**0.9328**	0.9908	**0.9328**

으로 식별 정확도가 높다는 것을 확인할 수 있었다. 특히, HSV의 경우 학습에 사용된 카테고리의 수가 증가할수록 식별 정확도가 향상되지만, 학습에 사용된 카테고리의 수가 작을 때도 학습에 사용되지 않은 카테고리에 포함된 영상에 대한 식별 정확도가 높은 것을 확인할 수 있었다. 즉 일반화 능력이 매우 우수한 것을 확인하였다. 한편, 카테고리 수가 7일 때 ("bridge" 카테고리가 추가됐을 때)를 보면, RGB와 HSV 색 공간을 사용할 때는 카테고리(또는 콘텐츠)가 증가했음에도 식별 정확도가 감소한다. 반면, YCbCr 색 공간을 사용할 때는 식별 정확도가 증가하는 것을 확인할 수 있다. 이로부터 콘텐츠의 종류나 유형에 따라 분류나 식별에 유용한 정보를 가진 색 공간이 다르다는 것을 추측할 수 있었다.

4.4 드롭아웃 적용 여부에 따른 위조 영상 식별 정확도

4.3절의 결과에서 색 공간에 따른 식별 정확도의 변화를 관측했을 때 일정한 변화를 특정하기 어려웠다. 따라서, 신경망 일반화의 대표적 기법인 드롭아웃을 모델 종단에 적용하여 학습 시 과적합되는 문제점을 보완하고자 하였다. 드롭아웃 비율은 일반적으로 많이 사용되는 0.2와 0.5를 사용하였다.

Table 4와 5는 드롭아웃 비율에 따른 식별 정확도를 보여주고, Fig. 3은 드롭아웃 사용 유무 및 드롭아웃 비율에 따라 학습에 포함되지 않은 데이터를 평가했을 때의 평균 식별 정확도가 어떻게 달라지는지를 그래프로 나타낸 것이다. 그림 3의 평균 정확도 추세선을 보면 드롭아웃이 적용되기 이전보다 적용되었을 때의 정확도가 전체적으로 높은 것을 확인할 수

Table 4. Fake image identification rates of Pelee+HPF with a dropout ratio of 0.2.

Train	Test											mean ACC (Trained)	mean ACC (Not Trained)
	cat	church	train	airplane	bus	cow	bridge	bedroom	classroom	restaurant	sheep		
Single object(cat)	1	**0.5**	**0.8592**	**0.9682**	**0.5242**	**0.9592**	**0.5005**	**1**	**0.5012**	**0.5**	**0.9252**	1	**0.7238**
Mix_2 (+ church)	0.9915	0.9968	**0.724**	**0.8175**	**0.5132**	**0.927**	**0.9268**	**0.8565**	**0.893**	**0.8365**	**0.9088**	0.9942	**0.8226**
Mix_3 (+ train)	0.9948	0.9992	0.9792	**0.9755**	**0.9232**	**0.9718**	**0.9408**	**0.9158**	**0.9465**	**0.9425**	**0.9688**	0.9911	**0.9481**
Mix_4 (+ airplane)	0.937	0.6588	0.7272	0.868	**0.5**	**0.6082**	**0.5572**	**0.556**	**0.5512**	**0.5668**	**0.5632**	0.7978	**0.5575**
Mix_5 (+ bus)	0.9675	0.985	0.9552	0.8902	0.9628	**0.9248**	**0.7908**	**0.889**	**0.9318**	**0.8142**	**0.8982**	0.9521	**0.8748**
Mix_6 (+ cow)	0.9972	0.9978	0.998	0.9965	0.996	0.9968	**0.964**	**0.9558**	**0.9565**	**0.9368**	**0.996**	0.997	**0.9618**
Mix_7 (+ bridge)	0.9578	0.9342	0.9285	0.9612	0.9242	0.9422	0.965	**0.9608**	**0.984**	**0.9695**	**0.9085**	0.9447	**0.9557**
Mix_8 (+ bedroom)	0.949	0.9152	0.9358	0.8585	0.8572	0.9668	0.9482	0.8938	**0.8715**	**0.7278**	**0.9398**	0.9156	**0.8464**
Mix_9 (+ classroom)	0.9848	0.9805	0.9808	0.9712	0.9765	0.9855	0.9738	0.987	0.9962	**0.9795**	**0.9478**	0.9818	**0.9636**
Mix_10 (+ restaurant)	0.9965	0.9975	0.9938	0.9945	0.9868	0.9982	0.9955	0.9975	0.9985	0.999	**0.9695**	0.9934	**0.9695**

Table 5. Fake image identification rates of Pelee+HPF with a dropout ratio of 0.5.

Train	Test											mean ACC (Trained)	mean ACC (Not Trained)
	cat	church	train	airplane	bus	cow	bridge	bedroom	classroom	restaurant	sheep		
Single object(cat)	0.9992	**0.5**	**0.9418**	**0.9705**	**0.5368**	**0.9795**	**0.5**	**0.5**	**0.5**	**0.5**	**0.9638**	0.9992	**0.6892**
Mix_2 (+ church)	0.9975	0.9988	**0.8178**	**0.968**	**0.5575**	**0.9018**	**0.985**	**0.96**	**0.968**	**0.8242**	**0.8115**	0.9982	**0.866**
Mix_3 (+ train)	0.9998	0.9995	0.9998	**0.9952**	**0.5755**	**0.9588**	**0.8372**	**0.765**	**0.7948**	**0.7065**	**0.9572**	0.9997	**0.8238**
Mix_4 (+ airplane)	0.9928	0.8952	0.9778	0.9475	**0.7602**	**0.9728**	**0.926**	**0.8675**	**0.91**	**0.8758**	**0.9842**	0.9533	**0.8995**
Mix_5 (+ bus)	0.9982	0.9988	0.999	0.9978	0.9968	**0.99**	**0.963**	**0.9252**	**0.967**	**0.9172**	**0.9758**	0.9981	**0.9564**
Mix_6 (+ cow)	0.9682	0.9142	0.9342	0.8968	0.8585	0.9718	**0.6598**	**0.7048**	**0.7248**	**0.6795**	**0.9065**	0.924	**0.7351**
Mix_7 (+ bridge)	0.9978	0.9992	0.9978	0.9948	0.9978	0.9995	0.9985	**0.9672**	**0.9848**	**0.9268**	**0.892**	0.9979	**0.9427**
Mix_8 (+ bedroom)	0.9825	0.9962	0.9755	0.9912	0.986	0.9765	0.994	0.999	**0.9992**	**0.9752**	**0.9522**	0.9876	**0.9755**
Mix_9 (+ classroom)	0.93	0.9388	0.9365	0.879	0.8835	0.9688	0.9425	0.9032	0.9185	**0.814**	**0.8748**	0.9223	**0.8444**
Mix_10 (+ restaurant)	0.9905	0.996	0.9915	0.9702	0.9642	0.9962	0.9898	0.9945	0.997	0.9972	**0.915**	0.9887	**0.915**

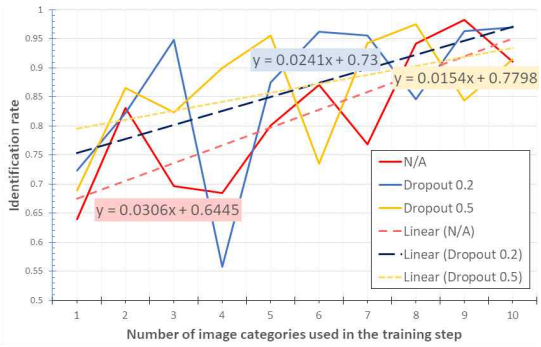

Fig. 3. Change in fake image identification rates for the images that are not included in the image categories used in the training step when using different dropout ratios.

있다. 특히 0.2의 경우 적용 전과 거의 평행하게 높다(추세선 참조)는 점에서 훈련 데이터에 포함되지 않은 데이터를 평가할 때 올바른 드롭아웃 값을 적용함으로써 딥러닝 모델의 강화된 일반화가 가능한 것을 알 수 있다.

4.5 일반화 능력 향상을 위한 최적 조합 및 Xception과의 성능 비교

앞 절들의 실험을 통해 Pelee+HPF 모델에 일반화 능력 향상에 효과적인 방법들을 조합하여 적용(즉, HSV 색 공간을 사용하고 드롭아웃 비율을 0.2로 설정)한 후 영상 카테고리 수를 추가하면서 RGB 색 공간을 사용하는 Xception 모델과 식별 정확도를 비교하는 실험을 진행하였다. Table 6, 7과 Fig. 4는 비교 결과를 보여주는데, 일반화 능력 향상에 효과적인 방법들을 조합한 경우, 오히려 성능이 떨어졌다. 즉, RGB 색 공간에서 드롭아웃을 사용할 때는 일반화 능력이 향상되었지만, HSV 색 공간에서는 드롭아웃을 사용하지 않는 것이 더 좋은 결과를 보여주었다. 드롭아웃을 사용할 경우, RGB와 HSV 색 공간에서

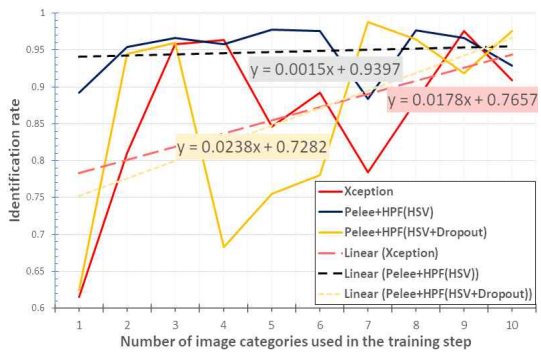

Fig. 4. Change in fake image identification rates of combinations of generalization methods, and comparison with Xception.

Table 6. Fake image identification rates of Pelee+HPF with the RGB to HSV color conversion and a dropout of 0.2.

Train	Test											mean ACC (Trained)	mean ACC (Not Trained)
	cat	church	train	airplane	bus	cow	bridge	bedroom	classroom	restaurant	sheep		
Single object(cat)	0.9822	**0.4852**	**0.7255**	**0.756**	**0.6105**	**0.6108**	**0.5537**	**0.6968**	**0.6188**	**0.5992**	**0.5948**	0.9822	**0.6251**
Mix_2 (+ church)	0.9105	0.9925	**0.968**	**0.9562**	**0.9685**	**0.9078**	**0.9735**	**0.9592**	**0.9838**	**0.911**	**0.8702**	0.9515	**0.9442**
Mix_3 (+ train)	0.9935	0.9942	0.9958	**0.976**	**0.954**	**0.9038**	**0.969**	**0.993**	**0.9975**	**0.9555**	**0.9302**	0.9945	**0.9599**
Mix_4 (+ airplane)	0.7368	0.7377	0.7148	0.6115	**0.7375**	**0.7405**	**0.744**	**0.551**	**0.6108**	**0.6122**	**0.782**	0.7002	**0.6826**
Mix_5 (+ bus)	0.855	0.5773	0.8828	0.7905	0.7892	**0.7645**	**0.6572**	**0.7102**	**0.8865**	**0.7185**	**0.793**	0.779	**0.755**
Mix_6 (+ cow)	0.8828	0.787	0.8298	0.908	0.6872	0.8272	**0.6948**	**0.881**	**0.8932**	**0.682**	**0.751**	0.8203	**0.7804**
Mix_7 (+ bridge)	0.9865	0.9957	0.9945	0.9958	0.9955	0.989	0.9958	**0.9938**	**0.9992**	**0.9835**	**0.9762**	0.9932	**0.9882**
Mix_8 (+ bedroom)	0.9898	0.9818	0.9932	0.9902	0.9668	0.985	0.9885	0.9905	**0.9928**	**0.962**	**0.938**	0.9857	**0.9643**
Mix_9 (+ classroom)	0.9565	0.9808	0.9845	0.9805	0.987	0.9442	0.9662	0.99	0.9888	**0.9638**	**0.8722**	0.9754	**0.918**
Mix_10 (+ restaurant)	0.9882	0.9972	0.9968	0.9962	0.995	0.99	0.9922	0.9975	0.9995	0.995	**0.9755**	0.9948	**0.9755**

Table 7. Fake image identification rates of Xception.

Train	Test											mean ACC (Trained)	mean ACC (Not Trained)
	cat	church	train	airplane	bus	cow	bridge	bedroom	classroom	restaurant	sheep		
Single object(cat)	0.9992	**0.4998**	**0.686**	**0.843**	**0.5197**	**0.8452**	**0.4995**	**0.4978**	**0.4998**	**0.4992**	**0.7592**	0.9992	**0.6149**
Mix_2 (+ church)	0.9992	0.999	**0.808**	**0.9652**	**0.585**	**0.874**	**0.905**	**0.7898**	**0.881**	**0.7002**	**0.786**	0.9991	**0.8105**
Mix_3 (+ train)	0.9988	1	0.9995	**0.9978**	**1**	**0.9995**	**0.9998**	**0.6638**	**1**	**0.9998**	**0.9995**	0.9994	**0.9575**
Mix_4 (+ airplane)	0.9995	1	0.9998	0.9988	**0.9995**	**0.9988**	**0.9998**	**0.7448**	**1**	**0.9998**	**0.9998**	0.9995	**0.9632**
Mix_5 (+ bus)	0.9965	0.9978	0.9985	0.9958	0.9968	**0.9812**	**0.7705**	**0.702**	**0.865**	**0.78**	**0.9775**	0.9971	**0.846**
Mix_6 (+ cow)	0.9988	0.998	0.9995	0.999	0.998	0.999	**0.9538**	**0.6642**	**0.9775**	**0.8715**	**0.9958**	0.9987	**0.8926**
Mix_7 (+ bridge)	0.9968	0.992	0.9962	0.9942	0.9952	0.9968	0.9915	**0.7165**	**0.7018**	**0.726**	**0.9925**	0.9947	**0.7842**
Mix_8 (+ bedroom)	0.9818	0.9835	0.9835	0.9745	0.9822	0.9898	0.9892	0.9788	**0.8965**	**0.768**	**0.979**	0.9829	**0.8812**
Mix_9 (+ classroom)	0.9658	0.9802	0.9518	0.9762	0.9795	0.975	0.9792	0.9932	0.9975	**0.996**	**0.956**	0.9776	**0.976**
Mix_10 (+ restaurant)	0.9832	0.958	0.9768	0.9878	0.9818	0.9802	0.967	0.9878	0.9972	0.9905	**0.909**	0.981	**0.909**

의 결과가 거의 동일했다(Fig. 3과 4 참조). 결과적으로, 본 논문에서는 일반화 능력을 향상시키기 위해서는, 입력 영상을 HSV 색 공간으로 변환하고 학습에 사용된 영상 카테고리 수를 늘리는 것이 가장 좋은 방법임을 확인하였다.

Xception 모델의 경우, 카테고리 수(또는 콘텐츠)가 증가함에 따라 식별 정확도가 향상되는 것은 마찬가지였으며, Pelee+HPF 모델에 드롭아웃을 적용한 것과 유사한 성능을 보였다. 그러나, HSV 색 공간을 사용하는 것보다는 식별 정확도나 일반화 능력이 떨어졌다.

5. 결 론

본 논문은 GAN에 의해 생성된 영상을 식별하는 것을 목적으로 위조 영상 식별을 위한 딥러닝 모델(Pelee)의 일반화 능력을 향상시키기 위한 방법들을 제시하고, 실험을 통해 성능을 검증하였다. 학습에 사용된 영상의 카테고리 수(또는 콘텐츠)가 증가할수록 학습에 사용되지 않은 카테고리(처음 보는 콘텐츠)에 대한 식별 정확도가 향상되었고, HSV나 YCbCr로의 색 공간 변환이나 드롭아웃을 적용함으로써 일반화 능력이 향상될 수 있음을 확인하였다. 그러나, 모든 방법을 함께 사용하는 것보다 HSV로의 색 공간 변환을 사용하고 학습에 사용되는 영상의 카테고리 수(또는 콘텐츠)를 늘리는 것이 일반화 능력 향상에 가장 효과적이었다. 즉, 색 공간 변환으로 인한 일반화 능력 개선 효과가 가장 월등했으며, 학습 데이터 증가를 통해 추가적인 개선 효과를 가졌으나, 색 공간을 변환하고 드롭아웃을 적용하는 것은 일반화 능력을 오히려 떨어뜨렸다.

본 논문에서는 기본 딥러닝 모델로 Pelee를 사용했지만, Xception을 비롯한 다른 딥러닝 모델의 일반화 능력을 향상시키기 위한 방법에 대한 추가 연구가 필요하다. 이는 딥러닝 모델에 따라 일반화 능력 향상에 도움이 되는 방법들이 달라질 수 있기 때문이다.

REFERENCE

[1] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and Improving the Image Quality of StyleGAN," *Proceeding of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8110-8119, 2020.

[2] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, and W. Shi, "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4681-4690, 2017.

[3] 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.

[4] H. Mo, B. Chen, and W. Luo, "Fake Faces Identification via Convolutional Neural Network," *Proceeding of the 6th ACM Workshop on Information Hiding and Multimedia Security*, pp. 43-47, 2018.

[5] N.T. Do, I.S. Na, and S.H. Kim, "Forensics Face Detection from GANs Using Convolutional Neural Network," *Proceeding of International Symposium on Information Technol-*

ogy Convergence, 2018.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, and Y. Bengio, "Generative Adversarial Nets," *Proceedings of the 27th International Conference on Neural Information Processing Systems*, Vol. 2, pp. 2672-2680, 2014.

[7] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive Growing of GANs for Improved Quality, Stability, and Variation," *arXiv preprint*, arXiv:1710.10196, 2017.

[8] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi, "Do GANs Leave Artificial Fingerprints?," *Proceeding of IEEE Conference on Multimedia Information Processing and Retrieval*, pp. 506-511, 2019.

[9] N. Yu, L.S. Davis, and M. Fritz, "Attributing Fake Images to GANs: Learning and Analyzing GAN fingerprints," *Proceeding of the IEEE/CVF International Conference on Computer Vision*, pp. 7556-7566, 2019.

[10] L. Verdoliva, "Media Forensics and Deepfakes: an Overview," *IEEE Journal of Selected Topics in Signal Processing*, Vol. 14, No. 5, 910-932, 2020.

[11] D. Cozzolino, J. Thies, A. Rossler, C. Riess, M. Nießner, and L. Verdoliva, "Forensictransfer: Weakly-Supervised Domain Adaptation for Forgery Detection," *arXiv preprint*, arXiv: 1812.02510, 2018.

[12] M. Du, S. Pentyala, Y. Li, and X. Hu, "Towards Generalizable Forgery Detection with Locality-Aware Autoencoder," *arXiv preprint*, arXiv: 1909.05999, 2019.

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

[14] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1251-1258, 2017.

[15] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, "LSUN: Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop," *arXiv preprint*, arXiv:1506.03365, 2015.

[16] R.J. Wang, X. Li, and C.X. Ling, "Pelee: A Real-Time Object Detection System on Mobile Devices," *arXiv preprint*, arXiv:1804.06882, 2018.

[17] S. Kang and H. Park, "Hierarchical CNN-Based Senary Classification of Steganographic Algorithms," *Journal of Korea Multimedia Society*, Vol. 24, No. 4, pp. 550-557, 2021.

이 정 한

2018년~현재 부경대학교 전자공학과 학사과정

관심분야 : 딥러닝, 컴퓨터비전

박 한 훈

2000년 한양대학교 전자통신전파공학과 졸업(공학사)

2002년 한양대학교 대학원 전자통신전파공학과 졸업(공학석사)

2007년 한양대학교 대학원 전자통신전파공학과 졸업(공학박사)

2008년~2011년 NHK방송기술연구소 박사후연구원

2012년~현재 부경대학교 전자공학과 교수

관심분야 : 증강현실, 인간컴퓨터상호작용, 3차원 영상처리/비전, 딥러닝 응용