KICSP Conference of Korea Institute of Convergence Signal Processing

위조 영상 식별을 위한 Xception 모델의 일반화 성능 분석

이정한, 박한훈*
부경대학교 전자공학과

Analysis of Generalization Perfomance of Xception Model for Fake Image Identification

Jeong-han Lee, Han-hoon Park*
Electronic Engineering, Pukyong University

Ⅰ. 연구 필요성 및 문제점

딥러닝(deep learning) 기술의 연구가 꾸준히 이뤄지면서 다양한 영상 생성 기법들이 제작되었다. 최근 적대적 생성 네트워크(GAN)[1]로 생성된 영상은 실(real) 영상과 육안으로 구분하기 힘들 정도로 우수한 화질을 가진다. 이로 인해 생성된 영상을 범죄에 악용하는 사회적 문제가 대두되고 있다. 따라서, GAN으로 생성된 영상을 식별하기 위한 딥러닝 기반 연구 또한 활발히 진행되고 있다. 최근 연구 결과를 보면, 육안으로 분류하기 힘든 위조(fake) 영상에 대해 매우 높은 식별 성능을 보여준다. 그러나, 학습에 사용된 영상과 다른 콘텐츠이거나, 학습에 사용된 영상과 다른 GAN 모델로 생성된 영상을 식별하는 성능은 크게 떨어진다는 문제점이 있다.

본 논문에서는 이러한 딥러닝 기반 방법들의 일반화 성능을 향상시키기 위해 학습에 사용되는 영상 카테고리 수를 점진적으로 늘리거나 보편적인 일반화 기법인 드롭아웃(dropout)을 적용해보고, Xception 모델[6]의 일반화 성능 변화를 실험적으로 분석한다.

Ⅱ. 실험 및 결과

2.1. 실험 환경

2.1.1. 데이터셋

학습을 위한 실 영상 데이터로 LSUN[2]의 카테고리를 임의로 11개(cat, church_outdoor, train, airplane, bus, cow, bridge, bedroom, classroom, restaurant, sheep)를 선정하였다. 위조 영상은 LSUN 데이터로 학습된 ProGAN[3]모델로 생성하였으며, 실 영상과 동일한 수를 사용하였다.

평가 데이터셋은 각 카테고리 별로 실 영상, 위조 영상 각각 2,000장으로 구성했다. ProGAN으로 생성된 “sheep” 카테고리 영상의 경우, 대표 평가 데이터셋으로 학습에 포함하지 않았다. 또한, StyleGAN2[4] 모델로 생성된 LSUN의 “horse” 카테고리 영상과 FFHQ[5] 영상으로 또 다른 평가 데이터셋을 구성하여 학습에 포함되지 않은 콘텐츠와 GAN 모델을 보유한 데이터에 대해서도 평가를 진행하였다.

2.1.2. 딥러닝 모델

본 논문에서 사용한 딥러닝 모델인 Xception[6]은 기존 모델인 Inception에서 수정된 Modified Depthwise Separable Convolution을 사용해 더 우수한 연산 성능을 확보한 모델이다. 우수한 성능에 비해 연산량이 적다는 장점도 가지고 있어 최근 위조 영상 식별 연구의 기본 백본 네트워크로 많이 활용되고 있다.

2.1.3. 실험 방법

먼저, 실 영상과 위조 영상을 각각 25,000장이 되도록 앞서 언급한 LSUN 카테고리의 영상을 동일한 비율로 맞춰 점진적으로 카테고리 수를 늘리면서 학습 데이터셋을 구성한다. 11개의 카테고리 별로 각각 구성된 평가 데이터셋에 대해 학습에 포함된 카테고리의 데이터셋과 포함되지 않은 데이터셋으로 나누어 평가를 진행해 정확도의 평균을 따로 구한다.

다음으로, 딥러닝 모델의 일반화 성능을 높일 수 있는 드롭아웃(dropout)을 Xception 모델의 마지막 층에 연결하여 정확도 변화를 확인한다. 드롭아웃 비율을 각각 0.2와 0.5로 설정하고, 첫 번째 실험과 동일한 과정을 반복해 평가하고 분석하였다.

마지막으로, 적용된 일반화 향상 방법들이 학습에 사용되지 않은 GAN 모델과 콘텐츠로 이뤄진 평가 데이터셋에 대해서도 유효한지 확인하였다. 이를 위해 가장 많은 카테고리(10개)를 사용하여 학습된 Xception 모델에 대해 드롭아웃의 유효성을 확인하였다.

2.2. 학습에 사용된 카테고리 수에 따른 정확도 분석 결과

카테고리 수를 점진적으로 증가하며 학습한 후, 학습에 사용되지 않은 카테고리의 위조 영상들에 대한 평균적인 식별 정확도 변화를 확인하였다(그림 1 참조). 카테고리 수에 따라 식별 정확도가 향상되는 것을 확인할 수 있으며, 이는 학습에 사용된 콘텐츠가 많아질수록 일반화 성능이 향상됨을 의미한다.

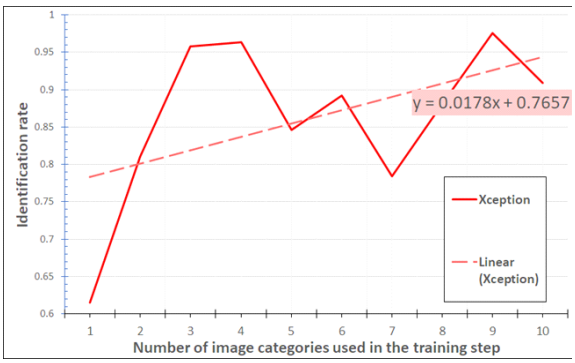

Fig. 1 Average accuracy and its linear trend line on evaluating unseen data according to changes in the number of image categories used in the training step

2.3. 드롭아웃 비율에 따른 정확도 분석 결과

그림 2에서 보는 것처럼, 드롭아웃을 적용함으로써, 대체적으로 일반화 성능이 향상되었으며, 특히, 드롭아웃 비율이 0.2일 경우, 학습에 사용된 카테고리 수에 상관없이 일관된 정확도 향상을 가져옴을 확인할 수 있었다.

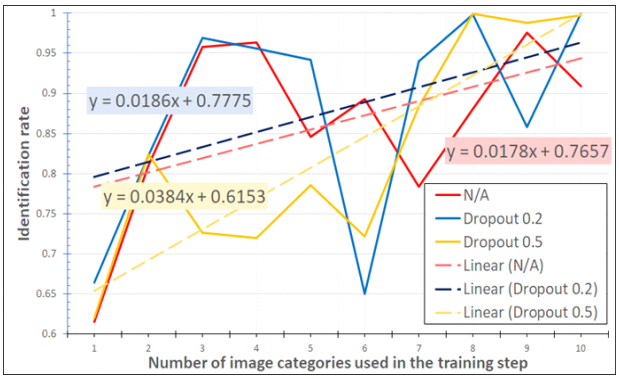

Fig. 2 Average accuracies and linear trend lines on evaluating unseen data according to dropout ratio

2.4. 학습 영상 생성에 사용된 GAN 모델과 다른 GAN 모델에 의해 생성된 위조 영상에 대한 정확도 분석 결과

그림 3은 10개의 카테고리 영상으로 학습된 Xception 모델의 StyleGAN2로 생성된 LSUN의 “horse” 카테고리 영상과 FFHQ 영상에 대한 정확도 결과를 보여준다. 우선, 드롭아웃을 적용하지 않을 경우, “horse” 위조 영상은 학습에 사용되지는 않았지만 LSUN에 포함된 영상이므로 유사한 콘텐츠를 포함하고 있는 것으로 보이며, 식별 정확도가 높았다. 그러나, FFHQ 위조 영상의 경우 식별이 전혀되지 않았다. 이는 영상 콘텐츠가 학습에 사용된 것과 완전히 다를 경우 식별이 불가능하다는 것을 의미한다.

반면, 드롭아웃을 적용함으로써, 전반적으로 식별 정확도가 크게 개선되었으며, FFHQ 영상에 대해서도 65% 이상의 정확도로 식별이 가능함을 확인하였다. 또한, 앞선 실험에서와 마찬가지로, 드롭아웃 비율이 0.2일 때, 성능 개선 효과가 높았다.

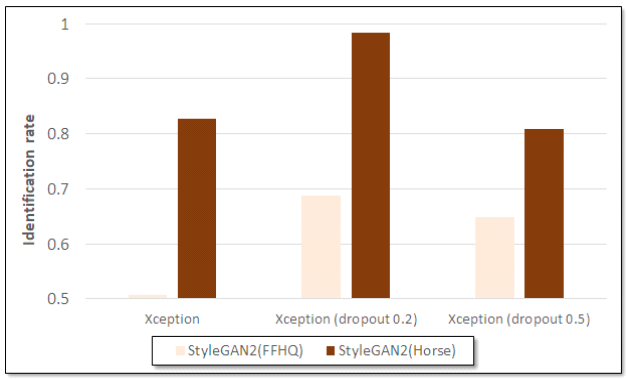

Fig. 3 Accuracy according to dropout ratio when evaluating fake images generated by StyleGAN2 model

Ⅲ. 결론 및 향후 연구

본 논문은 GAN으로 생성된 위조 영상을 식별하는 연구로서, 학습 시 보지 못한 콘텐츠를 포함하는 영상이나 학습 시 사용된 영상과 다른 GAN 모델로 생성된 영상에 대해 Xception 모델의 일반화 성능을 분석하고, 이를 향상시키기 위한 방법을 마련하기 위해 실험을 수행하였다. 학습에 사용된 영상 카테고리의 수가 증가할수록 새로운 콘텐츠를 가진 영상에 대한 식별의 정확도가 향상되는 것을 확인할 수 있었고, 드롭아웃을 통해 일반화 성능 보강이 가능하며 적절한 드롭아웃 비율이 주어졌을 때 성능이 보다 향상될 수 있음을 확인하였다.

그러나, 여전히 학습 시 사용된 영상과 다른 생성 모델로 생성되고, 학습 시 사용된 영상과 완전히 다른 콘텐츠를 가진 위조 영상에 대한 분류 정확도가 높지 못하기 때문에 이를 향상시키기 위한 추가 연구가 필요하다.

ACKNOWLEDGMENTS

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2021R1F1A1045749).

REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, and Y. Bengio, "Generative adversarial nets," Proc. of NIPS, 27, 2014.

[2] 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.

[3] 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.

[4] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," Proc. of CVPR, pp. 8110-8119, 2020.

[5] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," Proc. of CVPR, pp. 4401-4410, 2019.

[6] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," Proc. of CVPR, pp. 1251-1258, 2017.