패치 기반 딥페이크 영상 검출에 관한 연구

이정한, 박한훈*

부경대학교

jeonghan_lee@pukyong.ac.kr, *hanhoon.park@pknu.ac.kr

A Study on Patch-Wise Deepfake Image Detection

Jeonghan Lee, Hanhoon Park*

Pukyong National Univ.

요 약

본 논문은 딥페이크(deepfake) 영상 검출을 위해 영상을 패치로 나누고 패치별로 위조 여부를 판별한 후, 각 패치에 대한 판별 결과를 취합하는 방법을 제시하고, StyleGAN2로 생성된 영상에 대한 검출 성능을 실험을 통해 검증한다. 실험 결과, 패치 크기에 따라 검출 정확도는 달라졌으나, 전반적으로 패치 기반 방법은 검출 정확도를 크게 개선할 수 있으며, 신뢰도가 높은 패치를 선별하는 과정을 추가함으로써 검출 정확도를 보다 향상시킬 수 있음을 확인하였다.

Ⅰ. 서 론

본 논문에서는 기존 딥러닝 기반의 딥페이크(deepfake) 영상 검출 방법에서 입력 영상 전체를 한꺼번에 처리하던 것과 달리 입력 영상을 다양한 크기의 패치로 나누어 학습하고 분류한 후, 이로부터 보다 정확하게 딥페이크 영상을 검출할 수 있음을 실험적으로 검증하였다. 영상을 겹치지 않게 같은 크기의 패치로 나누어 학습 데이터셋을 구성하여 딥페이크 인식 모델을 학습하고, 입력 영상을 학습 때와 마찬가지로 동일한 크기의 겹치지 않는 패치로 나누거나 임의 위치에서의 패치를 선택하여 패치 기반 딥페이크 검출 성능을 분석하였다. 정확도 향상을 위해, 신뢰도가 높은 패치들만 선별하였으며 이를 통한 성능 변화를 분석하였다.

Ⅱ. 실험 및 결과

2.1. 데이터셋

본 논문에서는 실(real) 영상으로 고화질 얼굴 데이터셋인 FFHQ[1]의 영상을 임의추출하고, dlib[2]를 사용하여 얼굴 부위만 검출한 후 256x256 픽셀 크기로 조절하여 사용하였다. 위조(fake) 영상은 현재 가장 우수한 생성모델 중 하나인 StyleGAN2[3]를 사용하여 생성한 FFHQ 영상을 실 영상의 전처리와 동일하게 얼굴 부위만 추출하고 256x256 픽셀 크기로 조절하여 구성하였다. 실 영상과 위조 영상은 각각 학습 데이터 25,000장, 평가 데이터 2,000장씩 총 54,000장의 영상을 사용하였다. 패치로 구성된 데이터셋의 경우, 겹치지 않고 모든 픽셀을 사용하면서 패치 크기에 따른 정확도를 확인하기 위하여 모든 영상을 128x128, 64x64, 32x32, 16x16 픽셀 크기로 나누어 실험하였다.

2.2. 딥러닝 모델

본 논문에서는 Xception[4]을 사용하여 분류 예측값을 추출하였다. Xception은 실험을 통해 적은 연산량과 우수한 연산 성능으로 기존 딥페이크 영상 검출 연구에서 보편적으로 쓰이기 때문에 본 연구의 검출 모델로 채택되었다.

2.3. 실험 방법

실험1: 영상을 겹치지 않게 다양한 크기로 나눈 패치들을 각각 학습하고 동일한 크기의 평가 데이터셋 패치들을 단일 영상으로 고려하여 분류 정확도를 계산하였다.

실험2: 입력 영상 내 겹치지 않는 패치에 대해 real/fake 분류를 수행하고 real 또는 fake로 판별된 패치의 수로부터 입력 영상의 위조 여부를 결정하였으며, 4,000장의 입력 영상들에 대한 분류 정확도를 계산하였다. real과 fake로 판별된 패치의 수가 같을 경우, fake 영상으로 식별하도록 설정하였다.

실험3: 평가 데이터셋의 영상 내 임의의 패치를 추출하고, 임의의 패치들의 real/fake 분류 결과로부터 각 영상의 위조 여부를 판별하였다. 영상당 패치의 수는 real과 fake로 판별된 패치의 수가 같지 않도록 홀수로 구성하였다.

실험4: 영상으로부터 얻어진 패치들을 모두 사용하지 않고, real/fake 분류 신뢰도가 낮은 패치는 제외하고 분류 신뢰도가 높은 패치들만 선별하여 영상의 위조 여부를 판별하였다. 분류 신뢰도는 각 패치가 real 또는 fake일 확률의 차이가 클수록 높다고 판단하였다.

2.4. 성능 평가

딥러닝 모델은 교차 엔트로피 손실함수와 Adam 옵티마이저를 사용해 50 epoch 만큼 학습하였다. 성능 평가 지표로는 각 epoch 마다 정확도를 계산하여 가장 높은 정확도(top-1)와 가장 높은 정확도 5개의 평균(top-5), 그리고 전체 정확도의 평균(average) 이렇게 3개의 지표를 사용하였다.

2.5. 실험 1의 결과: 패치 크기에 따른 위조 패치 검출 정확도

표 1에서 보는 것처럼, 패치 크기에 따라 패치의 위조 여부 검출 정확도는 달라지지만, 큰 차이를 보이지는 않았다. 패치의 크기가 32x32일 때 정확도가 가장 높았다.

2.6. 실험 2와 3의 결과: 패치 크기에 따른 위조 영상 검출 정확도

표 2와 3에서 보는 것처럼, 패치 크기가 작을 경우 검출 정확도가 증가했으며, 이를 통해 패치 기반 위조 영상 검출이 유효함을 알 수 있다. 다만, 패치를 겹치지 않게 영상 전체에 걸쳐 얻을 때와 달리 임의의 위치에서 패치를 얻을 경우 영상 전체 정보를 활용할 수 없어 패치의 크기가 작지 않을 경우(즉, 64x64나 128x128인 경우) 검출 정확도가 크게 떨어졌다.

표 1. 패치 크기에 따른 위조 패치 검출 정확도

Patch Size	Top-1	Top-5	Average
256(Original)	96.58%	96.38%	94.80%
128	95.61%	95.49%	94.04%
64	92.73%	92.64%	91.46%
32	98.21%	98.08%	95.26%
16	96.61%	96.55%	93.34%

표 2. 패치 크기에 따른 겹치지 않는 패치 기반 위조 영상 검출 정확도

Patch Size	Top-1	Top-5	Average
128	97.90%	97.81%	96.51%
64	98.83%	98.79%	98.00%
32	100.00%	100.00%	98.45%
16	100.00%	100.00%	98.12%

표 3. 패치 크기에 따른 임의 위치의 패치 기반 위조 영상 검출 정확도

Patch Size	Top-1	Top-5	Average
128	54.80%	51.86%	50.41%
64	59.23%	58.66%	54.85%
32	100.00%	100.00%	98.43%
16	100.00%	100.00%	97.94%

표 4. 신뢰도가 높은 패치를 사용한 위조 영상 검출 정확도

Patch Size	Threshold	Top-1	Top-5	Average
128	θ = 1.0	97.90%	97.81%	96.51%
	θ = 0.9	98.62%	98.47%	97.32%
	θ = 0.8	98.83%	98.69%	97.64%
	θ = 0.7	99.22%	98.98%	97.63%
	θ = 0.6	99.33%	98.96%	97.53%
64	θ = 1.0	98.83%	98.79%	98.00%
	θ = 0.9	99.70%	99.64%	99.03%
	θ = 0.8	99.80%	99.73%	99.46%
	θ = 0.7	99.80%	99.74%	99.48%
	θ = 0.6	99.80%	99.74%	99.78%
32	θ = 1.0	100%	100%	98.45%
	θ = 0.9	100%	100%	98.71%
	θ = 0.8	100%	100%	98.65%
	θ = 0.7	100%	100%	98.64%
	θ = 0.6	100%	100%	98.64%
16	θ = 1.0	100%	100%	98.12%
	θ = 0.9	100%	100%	99.00%
	θ = 0.8	100%	100%	99.17%
	θ = 0.7	100%	100%	99.33%
	θ = 0.6	100%	100%	99.54%

2.7. 실험 4의 결과: 신뢰도가 높은 패치를 사용한 위조 영상 검출 정확도

표 4에서 θ는 신뢰도가 높은 패치를 선별하기 위한 문턱값으로, 각 패치에 대한 real과 fake일 확률의 차이를 크기에 따라 정렬한 후 상위 몇 %의 패치를 사용할 것인가를 결정한다. 예를 들어, 0.9는 상위 90%의 패치만을 사용한다. 결과적으로, 신뢰도가 높은 패치만을 선별하여 위조 영상을 식별함으로써 정확도가 개선됨을 알 수 있었다. 다만, θ가 작을수록 신뢰도가 매우 높은 패치만을 사용한다는 것인데, θ가 작을 경우 영상에 따라 문턱값을 만족하는 패치가 없을 수도 있으며, 이러한 영상은 실험 2에서와 같이 영상 내 모든 패치를 사용하여 위조 여부를 판별하였다. 결과적으로 θ를 작게 할수록 표 4의 결과는 표 2의 결과에 수렴하였다.

Ⅲ. 결론

본 논문에서는 StyleGAN2로 생성된 딥페이크 영상 검출을 위해 입력 영상을 패치로 나누고 각 패치의 real/fake 인식 결과로부터 영상의 위조 여부를 판별하는 방법의 성능을 실험적으로 검증하였다. 패치의 크기에 따라 위조 영상 검출 정확도는 달라졌으나, 패치 기반 위조 영상 검출 방법의 유효성은 확인할 수 있었다. 또한, 신뢰도가 높은 패치만을 선별함으로써 검출 정확도를 보다 개선할 수 있었다.

딥러닝 기반 딥페이크 영상 검출 방법은 학습 데이터셋과 평가 데이터셋의 생성 모델이 다르거나 콘텐츠가 상이한 경우 검출 정확도가 크게 떨어질 수 있는데, 향후 이러한 조건에서 패치 기반 검출 방법의 성능에 대한 분석을 수행하고자 한다.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) Grant by the Korean Government through the MSIT under Grant 2021R1F1A1045749.

참 고 문 헌

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

[2] Dlib, http://dlib.net/.

[3] 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.

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