생성된 비디오 탐지의 일반화 성능 향상을 위한 핵심 조건 분석*

이준희[O1], 김승재[2], 최승재[3], 이원규[4], 조명아[1†]
경희대학교 인공지능학과[1], 한신대학교 정보통신학부[2], 경희대학교 전자공학과[3], 경희대학교 우주과학과[4],
경희대학교 소프트웨어융합학과[1†]
jhlee39@khu.ac.kr[1], tmdwo8814@gmail.com[2], chltmd666@khu.ac.kr[3], fovert@khu.ac.kr[4],
maycho@khu.ac.kr[1†]

Analysis of Key Conditions for Generalizable Generated Video Detection

Junhee Lee[O1], Seungjae Kim[2], Seungjae Choi[3], Wongyu Lee[4], MyeongAh Cho[1†]
Department of Artificial Intelligence, Kyung Hee University[O1]
Department of Information and Communications, Han Shin University[2]
Department of Electronic Engineering, Kyung Hee University[3]
Department of Astronomy and space science, Kyung Hee University[4]
Department of Software Convergence, Kyung Hee University[1†]

요 약

최근 생성 AI의 빠른 발전으로 실제와 구분하기 어려울 정도의 품질을 갖춘 생성된 비디오가 인터넷에 많이 존재한다. 그런데 웹에 존재하는 수많은 데이터를 학습하는 대규모 생성 모델의 특성을 고려했을 때, 가짜 데이터를 함께 학습하는 경우 실제 분포가 손상되는 모델 붕괴 현상이 발생하게 되므로 사전에 실제와 구분하도록 해야 한다. 기존 생성 비디오 탐지 모델들은 학습 시 보지 못한 데이터의 소스 모델에 대해서는 잘 동작하지 못하였다. 이러한 문제 해결을 위해 본 논문에서는 실험과 관찰을 통해 생성된 비디오가 실제 비디오와 구분되어 갖는 특성을 명확하게 기술한다. 또한 이러한 사실이 탐지 모델의 일반화 성능 향상을 위해 사용되도록 하는 방향을 제시한다.

1. 서 론

최근 생성 AI 모델에 있어 빠른 발전으로 이미지뿐만 아니라 비디오 역시 실제 동영상과 구분하기 어려울 정도의 수준까지 다다랐다. 이러한 기술이 존재하기에 개개인은 창의적이고 개인화된 컨텐츠를 자유롭게 누릴 수 있다는 장점이 존재하면서도, 한편으로는 딥페이크를 남용하여 거짓 정보가 만들어지기도, 인간 창작물의 가치가 훼손되기도 하는 단점까지 동시에 존재한다.

그런데 이와 더불어 특히 생성 모델 연구 관련하여 중요한 사실이 존재하는데, 생성 모델을 학습하기 위해 실제가 아닌 생성된 데이터를 사용하게 되면 모델 붕괴 현상이 발생할 수 있다.[1] 이는 생성 모델이 실제 데이터의 분포만을 반영해야 하는 반면, 실제 데이터와는 다른 분포를 갖는 데이터가 함께 반영되어 점차 지식이 왜곡되는 현상이다. 현재도 실제와 유사한 생성 비디오들이 여러 소셜 미디어에 많이 업로드되고 있다는 점, 시간이 지날수록 AI 기술 사용자 수가 확대될 것이라는 점, 그리고 대규모 생성 모델 학습 시 인터넷 상에 존재하는 수많은 데이터를 무작위로 가져와 입력한다는 점을 고려했을 때 앞으로 발생 가능한 시나리오이다. 따라서 생성된 가짜 데이터를 탐지할 수 있는 방법이 필요하며, 이를 통해 학습 이전에 배제하도록 해야 한다.

* "본 연구는 과학기술정보통신부 및 정보통신기획평가원의 2024년도 SW중심대학사업의 결과로 수행되었음"(2023-0-00042)
† 교신저자

단순 생성된 데이터 탐지에 대한 연구는 이전부터 이어져 왔으며[2], 그중 비디오 도메인 관련 연구는 주로 딥페이크 탐지에 집중되어 있었다. 딥페이크는 비디오 상에서 특정 인물의 얼굴을 다른 얼굴로 대체하는 face swap, 얼굴 표정을 바꾸는 face reenactment가 가장 주요한 기능이므로, 탐지 시에도 사람의 특성인 facial landmark와 같은 정보를 단서로 사용하는 방법이 여럿 존재했다.[3] 하지만 모델이 생성해내는 비디오 컨텐츠의 다양성이 증가하면서 최근에는 딥페이크만이 아닌 생성된 모든 비디오를 식별하기 위한 모델들도 연구되는 추세이다. AIGVDET[4], DuB3D[5], 그리고 그 외에도 여러 방법이 제안되어 비교적 괜찮은 식별 성능을 달성했다. 하지만 모델 학습에 사용했던 비디오를 생성했던 모델 외에 다른 새로운(unseen) 모델로부터 생성된 비디오에 대해서는 현저히 낮은 성능을 보였다.

따라서 해당 논문에서는 이와 같이 공통적으로 일반화 성능이 저조하게 나타나는 현상에 대한 자세한 원인 분석을 다루고자 한다. 우선, (1) 앞선 연구들에서는 생성된 비디오에서 하나의 독립적인 프레임에도 실제와는 다른 생성기의 흔적이 남아 모델 학습에 있어 식별을 위한 단서 역할을 한다는 의견을 제시하였는데, 이러한 주장의 **타당성을 DIRE[6] 기법을 통해 검증**한다. (2) 그리고 모델 학습에서 사용했던 **생성된 비디오 데이터들이 갖는 공통적인 특징을 제시**하고, (3) 이러한 특징이 **일반화 성능 향상에 있어서 중요한 역할을 할 것임에 대한 타당한 근거를 수식과 그림을 통해 제시**한다. (4) 마지막으로 이전 연구들에서 데이터로 사용한 **생성 비디오의 FPS(Frames Per Second) 와 일**

반화 성능의 관계에 대해 역설하며 앞으로의 연구에 사용하면 더 도움 될 비디오 데이터의 특성을 제안한다.

해당 논문의 주된 기여는 다음과 같다.

- 생성된 비디오에서 독립적인 프레임이 실제 비디오와는 다른 흔적을 남겨 모델 성능 향상에 기여한다는 주장에 대한 타당성을 분석 실험을 통해 검증한다.
- 자세한 관찰을 통한 분석을 통해 생성 비디오 데이터들이 갖는 공통적인 특징을 보이고, 일반화 성능 향상을 위해 이러한 정보가 사용되어야 하는 이유를 제시한다.
- 앞의 분석 결과들을 통해 앞으로의 생성 비디오 탐지 모델 학습에 있어 사용을 지향해야 하는 비디오 데이터의 특성을 제안한다.

2. 관련 연구

2.1 DIRE(DIffusion Reconstruction Error)

DIRE[6]는 diffusion 모델에 의해 생성된 이미지를 탐지하기 위해 제안된 방법이다. 모델에 의해 생성된 이미지는 실제 이미지와 달리 생성 과정을 한 번 거쳤기 때문에 해당 이미지를 다시 복원하더라도 처음과 훨씬 유사할 것이라는 주장에서 고안되었다. 복원 전과 후의 두 이미지의 차이를 계산하는데, 식은 아래와 같다.

$$DIRE(\boldsymbol{x}_0) = |\boldsymbol{x}_0 - \boldsymbol{R}(\boldsymbol{I}(\boldsymbol{x}_0))|$$

$\boldsymbol{x}_0$는 실제, 혹은 diffusion에 의해 생성된 이미지이며, $\boldsymbol{I}$는 DDIM inversion[7]을 통해 가우시안 노이즈 이미지로 변환하는 과정, $\boldsymbol{R}$은 ADM(Ablated Diffusion Model)[8]을 통해 원본과 같아지도록 다시 복원하는 과정을 나타낸다. DIRE는 원본 이미지와 복원된 이미지의 차이를 이미지 자체에서 계산하기 때문에 DIRE 결과는 항상 원본 이미지와 동일한 크기를 갖게 된다. 위에서 언급했듯이, 생성된 이미지는 실제 이미지보다 비교적 다시 복원되었을 때 처음과 유사하기에 $DIRE(\boldsymbol{x}_0)$ 값이 상대적으로 더 작게 나오며, 이에 따라 시각화했을 경우 더 어두운 모습을 보인다.

3. 일반화 성능 저하 현상 원인 분석 및 제안 방법

3.1 단일 비디오 프레임의 DIRE

생성 비디오 탐지 모델 학습을 위해 비디오에서 랜덤하게 단일, 혹은 특정 개수의 프레임을 추출한 후 모델에 입력함으로써 해당 정보를 추론 단서 중 일부로 사용하는 방법론이 상당수 제안되었다. 이와 같은 방식을 적용할 경우 연속된 프레임을 샘플링한 것이 아니기 때문에 프레임 사이의 관계(temporal domain)는 고려되지 않고 오직 이미지 내 공간 축 상의 정보(spatial domain)만 고려되는 것이다.

하지만 이렇게 독립적인 프레임 내에 존재하는 정보가 생성 비디오 탐지에 있어서 정말로 유의미한지 검증하는 실험은 없었기에 본 논문에서 처음으로 타당성을 확인하고자 하였다. 그리고 검증 방법으로 위에서 설명한 DIRE 방법을 사용한다.

<표 1> 생성 비디오 탐지를 위한 검증 데이터셋 비교

Class	Model	Samples
Generated(T2V)	Lavie[8]	1,000
	VideoCrafter2[9]	1,000
	Sora[10]	48
Real	-	1,000

<표 1>은 검증을 위해 준비한 비디오 데이터셋의 전체 구조이다. SOTA 성능을 보이는 T2V(Text-to-Video) 모델에 해당하는 Lavie[8], VideoCrafter2[9], Sora[10] 세 모델로부터 생성된 비디오들과 가장 아래 행에 위치하는 실제 비디오 데이터셋(MSR-VTT[11])을 각각 표시된 개수만큼 사용했다. 각 비디오에서 하나의 프레임만을 샘플링한 후 해당 프레임의 DIRE 값을 계산했다. 이후 DIRE 이미지의 모든 픽셀에 대해 평균을 취하여 한 비디오 샘플 당 하나의 DIRE 값을 갖도록 했다. 마지막으로 각각의 동일한 생성 모델에 해당하는 모든 샘플 내에서 DIRE 값에 대한 평균을 구하여 <표 2>의 최종 결과를 도출하였다.

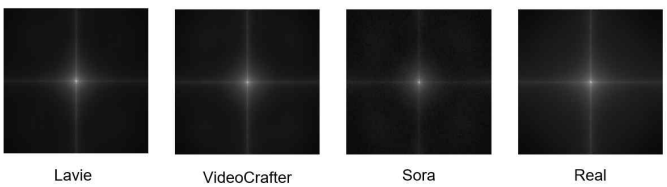
<표 2> 검증 데이터셋의 DIRE 결과

Class	Model	Avg DIRE
Generated(T2V)	Lavie[8]	23.76
	VideoCrafter2[9]	29.55
	Sora[10]	34.29
Real	-	23.71

비디오에서 샘플링한 단일 프레임이 생성 비디오 식별을 위한 유의미한 단서로 작용한다면 비록 DIRE가 diffusion 생성 이미지를 대상으로 측정한 매트릭일지라도, diffusion 기반 생성 모델인 Lavie와 VideoCrafter2에서 확연하게 DIRE Value가 작게 나와야 한다. 하지만 결과를 보니 오히려 실제 비디오에서 값이 비슷하거나 더 작게 나오는 경향을 보였다. 따라서 이와 같은 방법을 사용하여 모델에게 탐지를 위한 올바른 단서를 제공하지 못했기에 일반화 성능이 저조한 현상이 발생한다.

정량적 분석을 넘어서 결과 확인을 위해 직접 시각화도 진행하였는데, <그림 1>은 각 모델 샘플에 대한 DIRE 결과 중 임의로 하나씩 제시하며, <그림 2>는 각 생성 모델의 샘플에 대응하는 모든 DIRE 결과에 FFT(Fast Fourier Transform)를 통해 주파수 도메인으로 변환 후 평균을 취한 결과이다. 시각화를 통해 살펴본 결과 마찬가지로 실제 비디오와 구분되는 생성 흔적은 찾아볼 수 없다.

<그림 1> 검증 데이터셋의 DIRE 결과 시각화



<그림 2> DIRE 결과의 주파수 영역 시각화

3.2 생성된 비디오의 대표적 특성

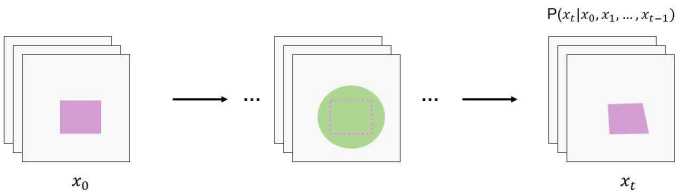
일반화 성능 향상을 도모하기 위해 생성 비디오가 실제 비디오와 구분되어 갖는 특징을 찾아내려 하였고, 이를 위해 수많은 데이터를 깊게 관찰한 결과 대부분의 데이터에서 다음과 같은 한 가지 공통점을 찾아낼 수 있었다.

<그림 3> 생성된 비디오의 대표적 특성 예시 1

<그림 4> 생성된 비디오의 대표적 특성 예시 2

<그림 3>과 <그림 4>는 Sora[10]를 통해 생성된 서로 다른 두 비디오에서 각각 프레임을 추출한 모습이다. <그림 3>에서 빨간 원을 보면 해당 인물이 가려지기 전과 후의 서 있는 방향이 바뀌어 있다. <그림 4>도 마찬가지로 표시한 부분이 가려지기 전에는 있던 물체가 가려진 후에 사라진 것을 볼 수 있다. 즉, 실제 비디오와 달리 생성된 비디오에서는 특정 물체가 가려진 후 일정 프레임 이후에 다시 영상에 나타났을 때 형상이 많이 바뀌는 특징이 존재한다.

이러한 현상은 비디오 프레임이 생성되는 원리에 의해 당연히 발생할 수밖에 없다. 비디오 프레임 생성 시 영상 내의 모션과 컨텐츠 모두의 일관성을 고려하기 위해 <그림 5>와 같이 이전 프레임들을 종합적으로 고려하게 된다. 하지만 특정 물체가 가려지기 전(x_0)과 후(x_t)의 두 프레임은 사이에 존재하는 프레임들에 의해 간격이 상대적으로 멀기 때문에 x_0에 대한 지식이 비교적 먼 프레임에 위치한 x_t까지 전해지는 과정에서 약해지게 된다. 따라서 이와 같은 현상이 발생하게 되며, 더 좋은 성능의 모델이 등장할지라도 생성 모델의 특성상 확률에 입각한 추론이라는 점을 벗어날 수는 없기에 일반화 성능 향상에 있어서 중요한 단서이다.



<그림 5> 프레임 생성 시 가려진 물체의 형상이 변환

3.3 모델 학습에 있어 지향해야 하는 데이터의 특성

앞서 두 분석을 통해 생성 비디오 탐지를 위해서는 프레임 각각이 아니라 시간 축을 기준으로 프레임 간의 일관성에 집중해야 한다는 사실을 알았다. 낮은 fps를 갖는 비디오 데이터를 사용하게 되면 프레임 사이의 변화가 커 사실상 프레임 각각을 사용하는 것에 가까워진다. 따라서 프레임 간 연결이 매끄러운 데이터로 학습하는 것이 일반화 성능 향상에 더 도움 될 것이다.

4. 결론 및 향후 연구

본 논문에서는 DIRE 결과와 생성된 비디오에 대한 자세한 관찰을 통해 생성 비디오 탐지의 일반화 성능 향상을 위해서는 독립적으로 각 프레임의 공간 축에 집중하는 것이 아니라, 비디오의 시간 축을 기준으로 프레임 간 일관성에 집중해야 함을 보였다. 이에 따라 앞으로의 모델 학습에 있어서 사용할 더 나은 데이터의 기준도 정의할 수 있었다.

그러나 좋은 성능을 위해서는 큰 규모의 데이터가 필요하며, 데이터의 개수를 크게 가져가면서 동시에 fps도 높이면 학습 시뿐만 아니라 데이터 구축 시에도 굉장히 비용이 많이 들게 된다. 따라서 적은 데이터로도 대규모 데이터를 사용한 것과 유사한 효과를 낼 수 있는 간접적인 방안에 대한 연구를 향후 진행할 것이다.

참고문헌

[1] Shumailov, Ilia, et al. "AI models collapse when trained on recursively generated data." Nature 631.8022 (2024): 755-759.

[2] Lin, Li, et al. "Detecting multimedia generated by large ai models: A survey." arXiv preprint arXiv:2402.00045 (2024).

[3] Yu, Peipeng, et al. "A survey on deepfake video detection." Iet Biometrics 10.6 (2021): 607-624.

[4] Bai, Jianfa, Man Lin, and Gang Cao. "AI-Generated Video Detection via Spatio-Temporal Anomaly Learning." arXiv preprint arXiv:2403.16638 (2024).

[5] Ji, Lichuan, et al. "Distinguish Any Fake Videos: Unleashing the Power of Large-scale Data and Motion Features." arXiv preprint arXiv:2405.15343 (2024).

[6] Wang, Zhendong, et al. "Dire for diffusion-generated image detection." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

[7] Song, Jiaming, Chenlin Meng, and Stefano Ermon. "Denoising diffusion implicit models." arXiv preprint arXiv:2010.02502 (2020).

[8] Wang, Yaohui, et al. "Lavie: High-quality video generation with cascaded latent diffusion models." arXiv preprint arXiv:2309.15103 (2023).

[9] Chen, Haoxin, et al. "Videocrafter2: Overcoming data limitations for high-quality video diffusion models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

[10] OpenAI, "Sora: Creating video from text." https://openai.com/sora, 2024.

[11] Xu, Jun, et al. "Msr-vtt: A large video description dataset for bridging video and language." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.