Diffusion 모델 기반 AI Inpainting 이미지의 탐지 및 조작 영역 판별에 관한 연구

남윤주 [1], 남윤수 [2], 김민규 [1], 이재호 [1*]
포항공과대학교 [1], Independent Researcher[2]

yoonjoo22@postech.ac.kr, yoon.nam82@gmail.com, minkyu4506@postech.ac.kr,
*jaeho.lee@postech.ac.kr

Detection and Localization of AI-Inpainted Images Using Diffusion Models

Yoonjoo Nam[1], Yoonsoo Nam[2], Minkyu Kim[1], Jaeho Lee[1*]
POSTECH[1], Independent Researcher[2]

요 약

최근 생성형 인공지능의 발달로 조작 이미지의 정교함이 증가하면서, AI 생성 inpainting 이미지 탐지가 점점 더 어려워지고 있다. 본 연구는 diffusion 모델 기반 접근을 활용하여, AI 생성 inpainting 이미지의 조작 여부를 판별하고 조작된 영역을 식별하는 통합적 탐지 프레임워크를 제안한다. 구체적으로, 이 프레임워크에서는 사전 학습된 diffusion 모델의 디노이징 과정을 통해 inpainting 이미지의 reconstruction 을 얻고, 이와 원본 이미지 간의 패치단위 표현형 차이를 분석하여 조작 탐지에 활용한다. 우리는 실험을 통해 제안된 방법이 다양한 방식으로 inpainting 된 이미지들에 대해 적용 가능하다는 것을 보였으며, 이미지 수준 및 픽셀 수준의 탐지 성능에서 유의미한 결과를 확인하였다.

Ⅰ. 서 론

생성형 인공지능 모델의 발전으로 쉽게 고품질 이미지를 생성할 수 있게 되면서 가짜 뉴스나 딥페이크 범죄 등 심각한 사회적 문제가 대두된다. 이러한 위험을 완화하고 생성형 AI 기술에 대한 대중의 신뢰를 형성하기 위해서는 이미지의 AI 생성 여부를 판별할 수 있는 기술이 필수적이다. 특히, 실제 이미지의 일부 영역만을 조작한 inpainting 이미지는 기존 생성 이미지 탐지 기법으로 식별하기 어려운 경우가 많다. 따라서 이러한 AI 생성 inpainting 이미지에 특화된 탐지 기술이 요구된다. 나아가, inpainting 이미지의 전체적인 판별 뿐만 아니라 이미지 내 각 영역의 진위 여부를 구분하는 것 또한 중요하다. 이를 위해서는 이미지 수준의 탐지와 함께, 픽셀 수준의 정밀한 조작 탐지 기술이 병행되어야 한다.

본 연구에서는 diffusion 기반 접근법을 활용하여, AI 생성 inpainting 이미지의 조작 여부 및 조작 영역의 식별 가능성을 실험적으로 확인하였다. 사전 학습된 diffusion model 로부터 추출한 정보를 활용해 학습된 분류기는, inpainting 이미지의 조작 여부 판별과 조작된 영역의 추정 모두에 활용될 수 있다.

Ⅱ. 본론

1. Diffusion Model 을 이용한 이미지 처리

최근 사용자 프롬프트를 잘 반영하는 고품질 이미지 생성 능력으로 인해 diffusion model 에 대한 연구가 활발히 진행되고 있다. diffusion model 은 무작위의 노이즈를 점진적으로 제거하는 방식으로 이미지를 생성하며 정방향(noising)과 역방향(denoising) 단계로 구성된다. 이 때 노이즈가 적용되는 표현 공간에 따라 image-level diffusion, latent space 기반 latent diffusion model 로 나뉜다. 이러한 diffusion model 은 대규모 이미지 데이터를 학습해 사실적인 이미지를 생성할 수 있으나 여전히 생성된 이미지의 품질과 일관성에서 개선이 필요하다.

본 연구에서는 사전 학습된 diffusion model 의 디노이징 과정을 활용하여, 입력 이미지로부터 조작 여부 판별에 유용한 정보를 추출할 수 있다고 가정하였다. 실제로, 선행 연구에서는 diffusion model 을 이용해 reconstruction error 를 계산하고 이를 기반으로 diffusion model 생성 이미지에 대한 탐지가 가능함이 보고되었다 [1]. 따라서 본 연구는 diffusion model 의 디노이징 특성을 활용해 조작 여부와 영역 정보를 내포한 표현형을 추출하고, 이를 기반으로 분류기 학습을 수행하는 구조를 설계하였다.

2. Inpainting 이미지의 분류 및 조작 영역 탐지

Inpainting 이미지의 조작된 영역을 효과적으로 탐지하기 위해서, 본 연구에서는 입력 이미지를 더 작은 패치(patch)단위로 분할하여 처리하는 방법을 사용하였다. 이는 기존 연구에서 제안된 CLIP 이미지 인코더를

활용한 조작 영역 탐지 기법 [2]과 유사한 접근이다. 이미지의 각 패치를 인코딩해 representation 으로 변환한 후, reconstruction 이미지와 원본 이미지 간의 차이를 분석함으로써 조작 여부를 추정한다. 이 과정에서 얻어진 logit 정보와 예측 마스크 이미지를 통해서 조작 영역 탐지에 활용된다.

실험에서는 RePaint [3], Latent Diffusion Model (LDM) [4]과 같은 diffusion 기반 inpainting 모델로 생성된 이미지를 사용하였다. 학습 및 검증에 사용된 이미지 데이터셋은 CelebA-HQ (256x256) 이미지 데이터셋을 4 가지 inpainting 방법으로 조작하여 생성된 Dolos 데이터셋을 사용했다 [5].

학습은 NVIDIA RTX 4090 환경에서 30 에폭(epoch) 동안 수행되었으며, 성능 평가는 In-Domain 데이터셋을 대상으로 진행되었다. 평가 지표로는 이미지 수준의 탐지 성능 측정을 위해 이진 정확도(ACC)를, 픽셀 수준의 조작 영역 탐지를 위해 IoU(intersection-over-union)를 사용하였다. 이때, ACC 는 임계값 0.5 를 기준으로 계산하였다. 비교 모델로는 동일한 데이터셋을 활용해 학습 및 평가를 진행한 DeCLIP [2]을 선택하였다. 비교 모델의 이미지 수준 탐지 성능은 공개된 코드를 통해 측정하였고, 픽셀 수준 탐지 성능은 논문에서 제공한 결과 중 LDM subset 으로 학습된 모델의 성능을 인용하였다.

실험 결과, 제안된 프레임워크는 이미지 수준의 탐지 정확도에서 LDM subset 에서는 기존 연구의 성능에 근접했으며, RePaint-p2-9k subset 에서는 기존 연구보다 크게 향상된 성능을 보였다. 반면, 픽셀 수준의 조작 영역 탐지에서는 기존 방법에 비해 낮은 성능을 보여 향후 개선이 필요함을 시사한다. 이러한 결과는 diffusion 기반 접근이 inpainting 이미지의 진위 여부 판별에 효과적으로 활용될 수 있음을 보여주며, 특히 이미지 수준의 탐지에서 유의미한 가능성을 확인할 수 있다.

Methods	LDM		Repaint-p2-9k	
	ACC	IoU	ACC	IoU
DeCLIP	0.522	0.491	0.516	0.437
Ours	0.519	0.221	0.839	0.211

표 1. Dolos dataset 에서의 성능 비교

Ⅲ. 결론

본 연구에서는 사전 학습된 diffusion 모델을 활용하여, AI 생성 inpainting 이미지의 조작 여부를 탐지하고 조작된 영역을 식별하는 방법론을 제안하고 그 가능성을 실험적으로 검증하였다. 제안한 방법은 diffusion 기반의 inpainting 이미지에 효과적으로 적용될 수 있음을 보였으며, 특히 이미지 수준의 판별에서는 기존 방법과 비교해 의미 있는 성능을 확인할 수 있었다.

한편, 픽셀 수준의 조작 영역 탐지에서는 상대적으로 낮은 성능을 보인다는 한계가 있다. 이는 reconstruction 이미지와의 차이를 비교하는 방식이나 예측 마스크를 얻는 모델 구조에 기인할 수 있다. 향후 연구에서는 이러한 한계를 보완하기 위한 모델 구조의 개선을 모색할 예정이며, 다양한 inpainting 기법에 대한 일반화 성능을 확보하는 방향으로 연구를 확장할 계획이다.

이를 통해 이미지 조작 탐지 기술의 실용적 가능성을 제시하며, 생성형 인공지능의 사회적 문제를 완화하고 이미지 콘텐츠의 신뢰성을 확보하는 데 기여할 수 있을 것으로 기대된다.

참 고 문 헌

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

[2] Smeu, Stefan, Elisabeta Oneata, and Dan Oneata. "DeCLIP: Decoding CLIP representations for deepfake localization." *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025.

[3] Lugmayr, Andreas, et al. "Repaint: Inpainting using denoising diffusion probabilistic models." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

[4] Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

[5] Ţânţaru, Dragoş-Constantin, Elisabeta Oneaţă , and Dan Oneaţă . "Weakly-supervised deepfake localization in diffusion-generated images." *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2024.