Proceeding of Conference on Knowledge Information Technology and Systems, Vol. x, No. x, pp. xx-xx, February 20xx
https://doi.org/10.34163/jkits.20xx.xx.x.xxx

Proceedings of Conference on Knowledge Information Technology and Systems

pISSN 1975-7700, eISSN 2734-0570

http://www.kkits.or.kr

AI-Generated Image Detection Using Frequency-Integrated Swin Tranformer

Sang-Hun Kim[1,+], **Jeong-Hoon Kim**[1,+], **Wonsik Jung**[1]
[1]*Department of Artificial Intelligence, Konyang University*
Sang-Hun Kim and Jeong-Hoon Kim contributed equally to this work.
***Corresponding author** Wonsik Jung, wsjung@konyang.ac.kr*

Abstract

With the rapid advancement of generative artificial intelligence (AI) technologies, the need for reliable methods to distinguish AI-generated images from real ones has become increasingly important. While prior research has primarily focused on spatial domain features extracted by convolutional neural networks (CNNs), recent studies suggest that frequency domain characteristics of generated images provide effective discriminative cues. We introduce a Swin Transformer-based model for detecting AI-generated images by integrating frequency domain information extracted via Fourier Transform.

Experiments on the CIFAKE dataset show that attention-based fusion strategies slightly improve detection accuracy. The proposed integration strategies outperform conventional CNN and Transformer models, demonstrating the effectiveness of incorporating frequency features within Transformer-based image classifiers.

Keywords
AI-generated Image Detection
Swin Transformer
Fourier Transform
Frequency information
Attention mechanism

Article history
Received: 11. Jun. 20xx
Revised: 28. Jul. 20xx
Accepted: 12. Oct. 20xx
Published: 31. Oct. 20xx

1. 서 론

최근 생성형 인공지능 (Artificial Intelligence, AI) 기술의 급격한 발전으로 이미지 생성 분야에서도 사람이 구별하기 어려울 정도로 고품질의 합성 이미지가 제작되고 있다. 이러한 기술은 다양한 긍정적인 활용 가능성을 제공하지만, 동시에 가짜 뉴스, 디지털 사기, 저작권 침해 등 사회적 문제로 이어질 위험성 또한 커지고 있다 [1]. 이에 따라 생성형 AI 기반 합성 이미지와 실제 이미지를 자동으로 구분할 수 있는 신뢰성 높은 판별 기술의 필요성이 점차 강조되고 있다.

https://doi.org/10.34163/jkits.20xx.xx.x.xxx

기존의 생성 이미지 탐지 연구는 주로 Convolutional Neural Network (CNN) 기반의 공간 도메인(spatial domain) 특성에 초점이 맞추어져 왔으나, 최근 연구들은 생성 과정에서 발생하는 주파수 도메인(frequency domain)의 비정상적인 패턴이 중요한 판별 단서가 될 수 있음을 보고하고 있다 [2,3]. 특히 생성 모델 특유의 고주파 성분 왜곡이나 스펙트럼 불균형은 공간 도메인에서는 쉽게 포착되지 않는 차별화된 특징으로 작용할 수 있다.

이에 본 연구에서는 이미지의 주파수 성분 분석을 통해 생성된 이미지 판별 성능을 향상시키는 방안을 탐색하고자 한다. 구체적으로, Swin Transformer [4]와 Fourier Transform 기반 주파수 정보를 다양한 방식으로 결합하여, Transformer의 전역적 문맥 이해 능력과 주파수 기반 미세 패턴 인지 능력을 동시에 활용하는 모델 구조를 설계하였다.

실험에는 생성형 AI 기반 합성 이미지 판별용 대표 데이터셋인 CIFAKE [5]를 활용하였으며, 이를 통해 주파수 정보 활용 여부 및 통합 방식에 따른 판별 성능 변화를 체계적으로 분석하고자 한다.

2. 관련 연구

AI 생성 이미지의 판별 기술은 최근 다양한 접근법을 통해 발전하고 있다. 기존 연구들은 CNN 기반 공간 도메인 특성을 활용한 판별 기법뿐 아니라, 주파수 도메인 정보의 활용 가능성에 주목하고 있다. Xi와 Chen [6]은 딥페이크 이미지 탐지에 Swin Transformer 기반 모델을 적용하고, 전처리 단계에서 Error Level Analysis (ELA) 기법을 활용하는 방법을 제안하였다. ELA는 JPEG 압축 과정에서 발생하는 압축 불일치 영역을 시각화하여 딥페이크 판별에 활용한다. 해당 연구에서는 Swin Transformer가 CNN 기반 모델보다 효과적임을 확인하였으나, ELA 기법은 비압축 이미지 또는 저품질 이미지에 취약하다는 한계가 있다.

또한, Mahara와 Rishe [7]는 AI 생성 이미지 탐지 분야의 최신 동향과 다양한 탐지 방법론을 종합적으로 정리하였다. 특히, 기존 연구들에서 제안된 주파수 영역 분석 기반 접근법이 생성 모델 고유의 주파수 특성을 활용하여 효과적인 판별 성능을 달성할 수 있음을 보고하였다.

본 연구는 이러한 선행 연구들의 결과를 참고하여, 주파수 도메인 정보를 Swin Transformer 기반 모델과 융합하는 탐지 방법을 설계하고자 한다.

3. 주파수 특성을 고려한 판별 모델

본 연구에서는 Swin Transformer 기반 판별 모델에 Fourier Transform을 통해 추출한 주파수 도메인 정보를 통합하여, 생성형 이미지 판별 성능 향상을 목표로 하는 모델을 설계하였다. 전체 구성은 1) 데이터셋 설명, 2) 주파수 정보 추출 과정, 3) 주파수 정보 통합 전략, 4) 실험 설계로 구성된다.

3.1 데이터셋 설명(Dataset Description)

실험을 위해 CIFAKE 데이터셋 [5]을 사용하였다. CIFAKE는 CIFAR-10 데이터셋의 10개 클래스에 대해 실제 이미지와 Stable Diffusion 기반 생성 이미지를 동일한 수량으로 제공하며, 총 12만 장으로 구성된다. 전체 데이터는 10:1:1 비율로 학습용, 검증용, 테스트용으로 분할하여 사용하였다.

모든 이미지는 224x224 크기로 리사이즈한 후, Swin Transformer 및 비교 방법들에 입력하였으며, 학습의 일반화 성능 향상을 위해 정규화 및 다양한 데이터 증강 기법을 적용하였다.

3.2 주파수 정보 추출(Frequency Domain

Information Extraction)

이미지의 주파수 정보는 2D Fourier Tranform (FFT)을 통해 추출하였다. 먼저 이미지를 흑백(1채널)으로 변환한 후, 2D FFT를 적용하여 amplitude spectrum을 계산하였다. 이후 스펙트럼 값을 로그 스케일로 변환하여 저주파 및 고주파 성분 간의 차이가 균형 있게 반영되도록 처리하였다.

모델 입력 형식에 맞추기 위해 변환된 1채널 이미지를 3채널 형태로 복제하였으며, 최종적으로 생성된 3채널 log-scaled amplitude spectrum 맵은 원본 이미지와 다양한 방식으로 통합하여 모델 학습에 활용하였다. 이를 통해 모델이 생성형 이미지의 고유한 주파수 특성을 효과적으로 학습할 수 있도록 구성하였다.

3.3 주파수 정보 통합 전략(Frequency Information Integration Strategy)

이전 장에서 추출한 주파수 도메인 정보를 Swin Transformer 기반 모델에 통합하기 위한 다양한 전략을 실험하였다. 주요 전략은 다음과 같다: 1) 원본 이미지만 사용(Swin-Tiny), 2) 주파수 정보만 사용(Swin-Tiny-FFT), 3) 원본 이미지와 주파수 정보를 각각 Swin-Tiny 기반 특징 추출 모듈을 통해 추출한 후, feature concatenation을 통해 단순 통합(Swin-Tiny w/ concat), 4) 3과 유사하게 특징을 추출한 후, attention을 통해 공간-주파수 정보 융합 특징 표현(Swin-Tiny w/ attention). 이를 통해 주파수 정보 활용 방식에 따른 생성형 이미지 판별 성능 변화를 체계적으로 분석하고자 하였다.

2.3 실험 설계(Experimental Design)

모든 실험은 동일한 Swin Transformer Tiny (Swin-Tiny) 특징 추출 모델을 기반으로 수행하였으며, 입력 해상도는 224x224로 고정하였다. 모델 가중치는 ImageNet 데이터셋에서 사전 학습된 Swin-Tiny 모델의 가중치를 초기값으로 활용한 후, 실험 데이터셋을 이용하여 fine-tuning을 진행하였다. 총 10 epoch 동안 AdamW Optimizer로 학습하였으며, 초기 학습률은 5e-5, 배치 크기는 32, 손실함수는 Binary Cross-Entropy를 사용하였다.

데이터 증강 기법으로는 확대, 좌우 반전, 정규화, 밝기 및 대비 조절을 적용하였으며, 모든 실험은 동일한 모델 구조와 조건에서 수행하여 공정한 성능 비교를 보장하였다.

비교 방법으로는 ResNet18 [8], MobileNetV3 [9], EfficientNet-B0 [10], MixNet [11], VGG11 [12], DEIT [13]를 사용하여 제안한 통합 전략과 성능을 비교하였다.

평가 지표로는 정확도(Accuracy) 및 F1-score를 사용하였으며, 동일한 학습 및 평가 조건에서 비교 방법 및 통합 전략 간 성능 차이를 정량적으로 비교하였다.

표 1. CIFAKE 데이터셋 내 생성된 이미지에 대한 탐지 결과
Table 1. Performance for AI-generated Image Detection on CIFAKE Dataset

Model	Train	Valid	Test	F1-score
	Accuracy			
ResNet18	0.9786	0.9764	0.9765	0.9764
MobileNetV3	0.9973	0.9843	0.9858	0.9858
EfficientNet-B0	0.9985	0.9841	0.9839	0.9839
MixNet	0.9975	0.9830	0.9820	0.9820
VGG11	<u>0.9994</u>	0.9847	0.9794	0.9792
DEIT	0.9955	0.9813	0.9833	0.9832
Swin-Tiny	0.9974	0.9879	<u>0.9901</u>	<u>0.9901</u>
Swin-Tiny (w/ concat)	**0.9996**	**0.9904**	0.9861	0.9861
Swin-Tiny (w/ attention)	0.9988	<u>0.9899</u>	**0.9912**	**0.9912**

https://doi.org/10.34163/jkits.20xx.xx.x.xxx

4. 실험 결과 및 통합 전략 효과 분석

이 장에서는 다양한 주파수 정보 통합 전략이 Swin Transformer 기반 모델의 생성형 이미지 판별 성능에 미치는 영향을 분석하였다. 또한, 기존 CNN 및 Transformer 기반 비교 모델들과의 성능 차이를 평가하였다.

<표 1>은 제안한 주파수 정보 통합 전략과 ResNet18, MobileNetV3, EfficientNet-B0, MixNet, VGG11, DEIT 기반 비교 방법들의 성능을 비교한 결과를 나타낸다. <표 1>에서 확인할 수 있듯이, 원본 이미지와 주파수 정보를 단순 결합(w/ concat)한 경우보다 attention 기반으로 통합한 전략(w/ attention)에서 성능이 가장 높게 나타났다. 이는 주파수 도메인 정보가 생성형 이미지와 실제 이미지 간의 미세한 차이를 효과적으로 반영하며, Swin Transformer와의 구조적 특성과 결합할 때 판별 성능을 더욱 향상시킬 수 있음을 보여준다.

또한, 대부분의 CNN 기반 모델들도 일정 수준 이상의 성능을 기록하였으나, Transformer 기반 모델인 DEIT와 본 연구에서 다룬 Swin Transformer 기반 모델에서 상대적으로 더 높은 성능을 보였다. 특히, 주파수 정보를 통합한 학습 전략(Swin-Tiny w/ concat, w/ attention)은 모든 평가 지표에서 가장 우수한 성능을 기록하였다. 이는 Transformer의 전역적 문맥 이해 능력과 주파수 기반 미세 패턴 인지 능력이 상호 보완적으로 작용함을 보여준다.

5. 결론

본 연구에서는 Swin Transformer 기반의 생성형 AI 이미지 판별 모델에 Fourier Transform 기반 주파수 도메인 정보를 다양한 방식으로 통합하여 판별 성능 향상을 확인하였다. 실험 결과, 원본 이미지와 주파수 정보를 attention 기반으로 통합한 전략에서 가장 우수한 성능을 보였다.

이는 Transformer의 전역적 문맥 이해 능력과 주파수 기반 미세 패턴 인지 능력이 상호 보완적으로 작용하여, 생성형 이미지 특유의 주파수 영역 특성을 효과적으로 반영한 결과로 해석된다. 또한, 제안하는 주파수 정보 통합 전략 방법은 다양한 기존 모델 대비 전반적으로 높은 판별 성능을 보여, 주파수 정보 통합의 효과성과 Transformer 기반 접근법의 유효성을 입증하였다.

향후 연구에서는 다양한 해상도 및 생성 모델에 대한 일반화 성능을 추가로 검증하고, Transformer 구조 내 주파수 정보 활용 방식의 고도화를 추진할 계획이다. 또한, CIFAKE 외 고해상도 생성 이미지 데이터셋에 대한 추가 실험을 통해 본 연구의 접근법의 범용성과 실용성을 더욱 심층적으로 평가하고자 한다.

References

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative Adversarial Nets,* Advances in Neural information Processing Systems, Vol. 27, 2014.

[2] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, *Leveraging Frequency Analysis for Deep Fake Image Recognition,* In International Conference on Machine Learning, PMLR, pp. 3247-3258, 2020.

[3] R. Durall, M. Keuper, and J. Keuper, *Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks Are Failing to Reproduce Spectral Distributions*, In Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, pp. 7890-7899, 2020.

[4] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, and B. Guo, *Swin Transformer: Hierarchical Vision Transformer Using*

https://doi.org/10.34163/jkits.20xx.xx.x.xxx

Shifted Windows, In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012-10022, 2021.

[5] J. J. Bird, and A. Lotfi, *CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images*, IEEE Access, Vol. 12, pp. 15642-15650, 2024.

[6] A. J. Xi, and E. Chen, *Classifying Deepfakes Using Swin Transformers*, arXiv preprint arXiv:2501.15656, 2025.

[7] A. Mahara, and N. Rishe, *Methods and Trends in Detecting Generated Images: A Comprehensive Review*, arXiv preprint arXiv:2502.15176, 2025.

[8] K. He, X. Zhang, S. Ren, and J. Sun, *Deep Residual Learning for Image Recognition*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778, 2016.

[9] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, *Searching for MobileNetV3,* Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1314-1324, 2019.

[10] M. Tan, and Q. Le, *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*, Proceedings of the 36th International Conference on Machine Learning, PMLR Vol. 97, pp. 6105-6114, 2019.

[11] M. Tan, and Q. V. Le, *MixConv: Mixed Depthwise Convolutional Kernels*, arXiv preprint arXiv:1907.09595, 2019.

[12] K. Simonyan, and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, arXiv preprint arXiv:1409.1556, 2014.

[13] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, *Training data-efficient image transformers and distillation through attention*, Proceedings of the 38th International Conference on Machine Learning, PMLR Vol. 139, pp. 10347-10357, 2021.

Swin Transformer 기반 주파수 정보 통합을 활용한 생성형 AI 이미지 판별

김상훈[1], 김정훈[1], 정원식[2]

[1]*건양대학교 인공지능학과 학사과정*

[2]*건양대학교 인공지능학과 조교수*

요 약

본 연구에서는 Swin Transformer 기반 생성형 이미지 판별 모델에 Fourier Transform 기반 주파수 정보를 다양한 방식으로 통합하는 전략을 제안하였다. CIFAKE 데이터셋을 활용한 실험 결과, attention 기반 통합 전략에서 가장 우수한 성능을 확인하였다.