딥페이크 탐지를 위한 지식 증류 기반 경량화 모델

이지수 [1], 김지우 [1], 김준화 [2]
건양대학교 의료인공지능학과 [1], 건양대학교 인공지능학과 [2]
dmsdud45122@naver.com, kimjiwoo410@gmail.com, junhwakim@konyang.ac.kr

Knowledge Distillation-Based Lightweight Model for Deepfake Detection

Jisu Lee[1] , Jiwoo Kim[1], Junhwa Kim[2]
Department of Medical Artificial Intelligence, Konyang University[1]
Department of Artificial Intelligence, Konyang University[2]

요약

딥페이크 기술은 허위 콘텐츠의 빠르고 광범위한 확산으로 인해 사회적 문제가 야기되며 이에 따라 신속하고 정확한 탐지 기술이 요구된다. 본 연구는 지식 증류 기법으로 경량화된 소형 모델을 통해 딥페이크 이미지 악용으로 인한 피해를 줄이는 데 기여하고자 한다. 그 결과 기존 Teacher 모델 대비 FPS가 약 3배 증가하며 실시간 응용 가능성을 입증하였고 모델 파라미터 수는 99.73% 감소하였음에도 불구하고 테스트 정확도 94.29%를 기록하며 높은 성능을 보였다.

Abstract

Deepfake technology can rapidly disseminate false content, posing significant social risks and necessitating prompt, accurate detection. This study employs a knowledge-distilled, lightweight model to mitigate deepfake image misuse. Results show a threefold increase in FPS compared to the teacher model, validating real-time applicability. Additionally, parameters are reduced by 99.73%, yet the model achieves a 94.29% test accuracy, indicating strong performance.

1. 서 론

딥페이크(Deepfake)는 인공지능(AI)을 이용해 이미지, 음성, 영상 등의 데이터를 조작하여 실제와 유사한 콘텐츠를 만들어내는 기술 [1]이다. 초기에는 창의적 콘텐츠 제작을 위한 목적으로 개발되었으나, 최근 AI 기술의 비약적 발전과 딥페이크 제작 프로그램이 오픈소스로 공개되면서 누구나 쉽게 가짜 이미지, 음성, 영상을 만들 수 있게 되었다. 이러한 기술의 확산은 사회적, 윤리적 문제를 야기하고 있다 [2].

경찰청 통계에 따르면 허위 생산물 등 범죄 관련 발생 건수가 2021년 156건에서 2024년 7월 기준 297건으로 해마다 급증하고 있다 [3]. 특히 딥페이크 기술을 악용한 성 착취물 제작이 주요 문제로 부각되고 있으며, 이러한 영상은 SNS와 모바일 앱 등을 통해 빠르게 유포된다. 유포된 영상은 완전한 삭제가 어려워 2차, 3차 피해로 이어질 가능성이 크다.

딥페이크 탐지 기술이 꾸준히 발전하고 있지만, 딥페이크 생성 기술이 이를 앞서나가면서 기존 탐지 기술로는 최신 딥페이크 콘텐츠를 효과적으로 차단하기에 한계가 있다[4]. 특히, 딥페이크 콘텐츠는 유포 속도가 매우 빠르기 때문에 탐지 및 대응 과정에서의 지연이 피해를 확산시키는 주요 원인으로 작용한다 [5]. 따라서 경량화된 모델로 실시간으로 탐지하며 높은 성능을 유지하는 것은 딥페이크 확산을 방지하고 피해를 최소화하기 위해 필수적인 해결책으로 요구된다.

본 연구는 딥페이크 얼굴 이미지 데이터를 활용하여 지식 증류(Knowledge Distillation) [6]기법을 통해 딥페이크 이미지를 실시간으로 구분하는 방법을 제안한다. 이를 통해 높은 성능을 유지하면서도 경량화된 소형 모델을 개발하여 딥페이크 이미지 악용으로 인한 피해를 줄이는 데 기여하고자 한다.

2. 연구 방법

2.1 데이터 전처리

본 연구에서는 Kaggle에서 제공한 Deepfake-dataset (140k+dataset real or fake) [7] 데이터셋을 사용하였다. 데이터셋의 전처리 과정으로 모든 데이터 크기를 동일하게 하기 위해 이미지를 256x256 픽셀로 크기 조정(Resize)한 후 주변의 불필요한 정보를 제거하고 주요 객체인 얼굴에 집중하도록 CenterCrop을 사용하여 이미지의 중심에서 224x224 크기로 잘랐다.

전처리된 이미지를 모델에 입력할 수 있도록 텐서 형식으로 변환하고, 픽셀 값은 [0, 1] 범위로 정규화하였다. 마지막으로 ImageNet 데이터셋에서 계산된

평균값 '[0.485, 0.456, 0.406]'과 표준편차 값 '[0.229, 0.224, 0.225]'을 사용하여 이미지를 정규화하였다. 이러한 전처리는 모델의 안정적인 학습을 지원하며, 데이터 분포의 일관성을 유지하여 학습 속도를 향상시키는 데 기여한다 [8].

2.2 지식 증류 기법

본 연구에서 사용한 지식 증류는 아래의 그림 1과 같은 구조로 구성되어 있다. Teacher 모델의 Soft label과 Ground truth의 Hard label을 결합하여 Student 모델을 학습한다. 이를 통해 Student 모델이 Teacher 모델의 출력 분포를 학습하고 True labels(정답)에 따라 올바르게 동작하도록 한다. 또한 Teacher 모델의 Soft label은 학습 초기 단계에서 Teacher 모델 출력값을 미리 계산하여 저장한 Precomputed Outputs [9]를 활용하였다. 이러한 방법을 통해 학습 효율성을 향상시키고 계산 리소스를 절약하였다.

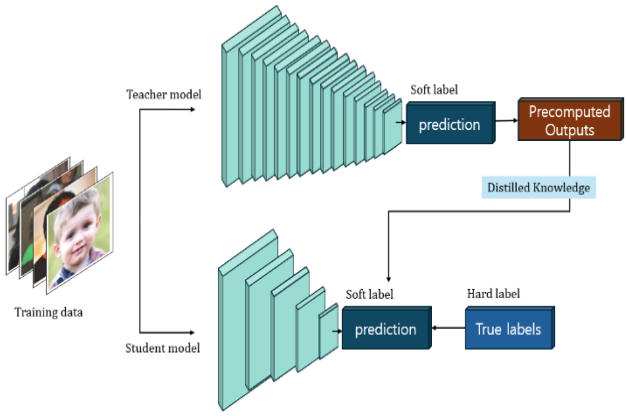

그림 1. 지식 증류 구조

본 연구에서는 지식 증류를 위해 두 가지 손실 함수를 사용하였다. Teacher 모델과 Student 모델 간의 출력 분포를 맞추기 위해 Soft Loss를 적용하고 True labels와의 차이를 줄이기 위해 Hard Loss를 추가하였다. 최종 손실 함수는 수식 (1)과 (2)로 정의되며, α 는 두 손실 간의 가중치를 조절하는 하이퍼파라미터이다.

$$L = \alpha \cdot KL(T,S) + (1-\alpha) \cdot L_{hard} \quad (1)$$

$$L = \alpha \cdot MSE(T,S) + (1-\alpha) \cdot L_{hard} \quad (2)$$

수식 (1)은 KL Divergence 기반으로 Soft Loss를 정의하며, Teacher 모델과 Student 모델 간의 출력 분포를 정렬하는 데 사용된다. 이를 위해 Teacher 모델의 출력을 Softmax와 온도 매개변수(temperature)를 적용하여 확률 분포를 생성하고 Student 모델의 Softmax 출력 분포와 비교한다. 온도 매개변수는 확률 분포를 부드럽게 만들어 Teacher 모델의 출력을 더 명확히 학습할 수 있도록 돕는다. KL Divergence는 이 두 분포 간의 차이를 측정하며, Teacher 모델의 지식을 Student 모델로 효과적으로 전달할 수 있도록 돕는다.

수식 (2)는 MSE(Mean Squared Error) 기반으로 Soft Loss를 정의하며, Teacher모델과 Student 모델 간의 출력 값 차이를 직접적으로 최소화하는 데 초점을 맞춘다. 이는 Teacher와 Student 간의 값 자체를 정렬함으로써 더 직관적인 방식으로 지식을 전달할 수 있도록 한다 [10].

2.3 Teacher 및 Student 모델

먼저, Teacher 모델을 선정하기 위해 높은 성능을 보이는 VGG19 [11], ResNet152 [12], EfficientNet-b7 [13]과 같은 대형 CNN 기반 사전 학습 모델을 사용하였다. 이러한 모델들은 수백 개의 레이어와 수백만 개의 파라미터를 가지고 있어 딥페이크 이미지의 복잡한 특징을 학습하는 데 효과적이다. 본 연구에서는 대형 모델의 학습 결과를 활용하여, Student 모델이 효과적으로 지식을 전달받도록 설계되었다.

Student 모델은 ResNet 구조를 기반으로 경량화된 소형 모델인 ResNet8을 직접 설계하였다. 이는 3채널 이미지를 입력으로 받으며 Conv2d 레이어를 사용해 16개의 출력 채널로 변환한 후, 각 BasicBlock은 두 개의 3x3 Convolution, Batch Normalization, ReLU 활성화 함수로 구성된다. Shortcut connection을 통해 입력과 출력을 더함으로써 네트워크가 깊어져도 기울기 소실 문제를 방지한다. 총 28개의 레이어가 쌓여 있으며, 잔차 학습을 통해 모델의 학습 성능을 개선하고자 한다. 이 모델은 계산 효율성을 높이며, 실제 응용 환경에서도 빠르게 동작할 수 있도록 설계되었다.

3. 실험 결과

3.1 딥페이크 데이터 세트

본 연구에서 사용한 데이터셋은 총 331,335개의 이미지 데이터로 구성되어 있으며, 해당 데이터셋은 그림 2와 같이 'Real', 'Fake' 두 가지 레이블로 분류된다. 'Real' 데이터는 실제 사람의 얼굴 이미지를 포함하고, 'Fake' 데이터는 딥페이크 기술로 생성된 가짜 얼굴 이미지를 포함한다. 두 레이블은 균형 잡힌 비율로 구성되어 있으며, 데이터셋에는 다양한 연령대와 여러 각도의 얼굴 이미지가 포함되어 있다. 데이터는 학습, 검증, 테스트 데이터로 각각 7:2:1의 비율로 분할하여 사용하였다.

그림 2. 데이터셋 예시

3.2 Teacher 모델 성능 평가

Teacher 모델로는 VGG19, ResNet152, EfficientNet-b7을 대상으로 하이퍼파라미터를 조정하여 성능을 비교하였다. 성능 평가는 Accuracy, F1-score, FPS(Frames Per Second)를 기준으로 진행하였다. FPS는 1초마다 처리한 이미지 수를 나타내며, 숫자가 클수록 추론 속도가 빠르다는 뜻이다[14]. 실험 결과, 표 1과 같이 EfficientNet-b7이 테스트 정확도 96.87%와 FPS 204.13으로 최종 Teacher 모델로 선정되었다. 최종 Teacher 모델로 선정된 EfficientNet-b7의 학습에 사용된 하이퍼파라미터는 표 2와 같다.

표 1. Teacher 모델 성능 비교

	VGG19	ResNet152	EfficientNet-b7
Accuracy(%)	96.35	96.48	**96.87**
F1-score(%)	96.31	96.47	**96.81**
FPS	327.71	239.57	**204.13**
Params(M)	33.0	58.1	**65.1**

표 2. EfficientNet-b7 최종 하이퍼파라미터

Epoch	20
Learning_Rate	1e-5
Batch_Size	16
Dropout	0.3
fc_hidden_dim	512

3.3 Student 모델 성능 및 지식 증류 효과

ResNet을 기반으로 설계된 Student 모델(ResNet8)은 파라미터 수가 0.145M로 Teacher 모델 대비 99.73% 감소하였다. 초기 실험은 ResNet8 기반으로 수식 (1)을 활용해 학습을 진행하였으며 ResNet8의 FPS는 617.72로 동일하다.

지식 증류를 적용하지 않은 ResNet8 모델과 지식 증류를 적용한 ResNet8 모델을 동일한 파라미터 조건(Epoch 20, Batch_Size 16, Learning Rate 1e-5, α 0.5)에서 학습한 후 성능을 비교한 결과는 표 3과 같다. 실험 결과, 지식 증류를 적용한 ResNet8 모델은 Accuracy에서 약 2%의 성능 향상을 보였으며, 특히 FPS에서 약 3배 높은 결과를 보인다. 이러한 결과는 Teacher 모델의 지식을 효과적으로 전달받았음을 보여주며, 본 연구에서 지식 증류를 활용한 실험이 효과적임을 입증한다.

표 3. ResNet8 모델의 지식 증류 적용 전후 성능 비교

	ResNet8 (no KD)	ResNet8 (with KD)
Accuracy(%)	86.70	**88.63**
F1-score(%)	87.31	**88.35**

이후 지식 증류를 적용한 ResNet8 의 성능을 Teacher 모델의 성능과 근접하도록 추가적인 실험을 진행하였다. 에포크는 20 으로 동일하게 설정하였으며, 그 결과 표 4 와 같이 나타났다. 배치 사이즈는 32, 학습률은 0.0001 로 설정하였다. 학습 스케줄러로는 ReduceLROnPlateau[15]를 사용하였으며, 이는 검증 데이터 세트에서 모델의 성능에 따라 학습률을 동적으로 조정하는 역할을 한다. 검증 손실이 3 회 연속으로 개선되지 않을 경우 학습률이 0.1 배 줄어들도록 설정하여 활용하였다. 또한 α를 0.4 로 설정했을 때 테스트 정확도가 93.92%로 가장 높은 성능을 보인다.

표 4. 하이퍼파라미터에 따른 Student 모델 성능 비교

Batch_Size	32	32	32	32
Learning Rate	0.00001	0.00001	0.0001	0.0001
Scheduler	-	-	-	ReduceLR OnPlateau
α	0.5	0.6	0.5	0.4
Accuracy(%)	92.36	92.32	92.74	93.92
F1-score(%)	92.05	91.99	92.51	93.85

하이퍼파라미터 조정만으로 최적의 성능을 내기에 한계가 있어 최적의 하이퍼파라미터는 유지한 채 수식 (1)에서 수식 (2)로 지식 증류의 손실 함수만 변경하여 실험을 진행하였다.

그 결과, 표 5 와 같은 성능을 보이며 최종적으로 기존 Student 모델만 사용했을 때에 비해 정확도가 7.5% 향상되고 FPS 도 증가한 것을 확인할 수 있다. Teacher 모델과의 성능은 2.58%의 차이로 거의 근접한 수치이다. 기존 Teacher 모델의 FPS 는 204.13 이었으나, 지식 증류를 적용한 후 617.72 로 약 3 배 증가하였다. 이는 1 초에 처리하는 수가 3 배

더 많아졌다는 것을 의미하며, 모델 파라미터 수도 Teacher 모델 대비 99.73% 감소하여 경량화된 소형 모델에서도 높은 성능을 유지할 수 있음을 보여준다.

표 5. 지식 증류 적용 후 최종 성능

	KL Divergence	MSE
Accuracy(%)	93.92	**94.29**
F1-score(%)	93.85	**94.26**
FPS	617.72	
Params(M)	0.145M	

4. 결론

본 연구는 딥페이크 얼굴 이미지를 활용하여 지식 증류 기법을 통해 딥페이크 이미지를 실시간으로 구분하는 방법을 제안하였다. Teacher 모델과 Student 모델 간의 FPS 와 파라미터 수를 비교하며 지식 증류를 진행하였다. 최종적으로 Teacher 모델은 EfficientNet-b7 을 선정하였고 Student 모델은 ResNet8 의 구조를 기반으로 모델을 직접 설계하였다. 지식 증류 기법을 적용한 결과, 기존 Teacher 모델 대비 FPS 가 약 3 배 증가하며 실시간 응용 가능성을 입증하였다. 모델 파라미터 수는 99.73% 감소하였음에도 최종적으로 테스트 정확도 94.29%를 기록하며 높은 성능을 유지하였다. 이를 통해 지식 증류 기법이 딥페이크 탐지 모델의 경량화 및 실시간 성능 개선에 효과적임을 확인하였다.

ACKNOWLEDGEMENT

본 과제(결과물)는 2024년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 지자체-대학 협력기반 지역혁신 사업의 결과입니다. (2021RIS-004)

본 연구는 2025년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업 지원을 받아 수행되었음(2024-0-00047)

참고문헌

[1] 최창욱, and 정유미. "국내외 언론이 바라본 딥페이크 기술." *Journal of Digital Contents Society* 23.5 (2022): 893-904.

[2]https://monthly.chosun.com/client/news/viw.asp?ctcd=C&nNewsNumb=202412100039

[3]https://www.ccnnews.co.kr/news/articleView.html?idxno=349297

[4]https://eiec.kdi.re.kr/publish/naraView.do?fcode=00002000040000100009&cidx=14938&sel_year=2024&sel_month=11&pp=20&pg=1

[5]https://news.mt.co.kr/mtview.php?no=2024091319472573195

[6] Hinton, Geoffrey. "Distilling the Knowledge in a Neural Network." *arXiv preprint arXiv:1503.02531* (2015).

[7]https://www.kaggle.com/datasets/tusharpadhy/deepfake-dataset

[8] Ioffe, Sergey. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." *arXiv preprint arXiv:1502.03167* (2015).

[9] Li, Chuan, and Michael Wand. "Precomputed real-time texture synthesis with markovian generative adversarial networks." *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*. Springer International Publishing, 2016.

[10] Kim, Taehyeon, et al. "Comparing kullback-leibler divergence and mean squared error loss in knowledge distillation." *arXiv preprint arXiv:2105.08919* (2021).

[11] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).

[12] He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[13] Tan, Mingxing, and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." *International conference on machine learning*. PMLR, 2019.

[14] 정승원. YOLOv8 기반 실시간 위암 검출 모델의 경량화 연구, 가천대학교 대학원 석사학위논문, 1-36, 2023

[15]https://pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.ReduceLROnPlateau.html