KAICTS Conference of Korea Artificial-Intelligence Convergence Technology Society

음성 데이터에서 폭력 감지를 위한 1D-2D 비교 및 모델 최적화

김지우[1†], 이신화[1†], 이지수[1†], 김준화[2]
[1]건양대학교 의료인공지능학과
[2]건양대학교 인공지능학과
† 위 저자는 본 논문에 동등하게 기여함.

Ⅰ. 서론

2023년 검찰청 통계에 따르면 폭력 사범으로 접수된 건수는 약 22만 건에 이르렀다. 현대 사회에서 일상 속 폭력 상황이 빈번하게 발생하면서, 이러한 위험을 실시간으로 탐지하는 기술에 대한 필요성이 증가하고 있다 [1]. 최근 이미지나 비디오 기반 폭력 탐지 연구가 활발하게 이루어지고 있지만, 이러한 방식은 집이나 CCTV의 사각지대에서 폭력을 감지하는 데 한계가 있다. 반면, 오디오는 이러한 시각적 제약을 극복할 수 있는 장점이 있다.

본 연구는 오디오 데이터를 활용하여 Residual Network [2] 기반의 딥러닝 모델을 통해 일상생활에서 발생하는 소리를 분석하고 폭력적 소리와 비폭력적 소리를 실시간으로 구분하는 방법을 제안하며, 스펙트로그램 변환 및 데이터 증강 기법을 적용함으로써 다양한 상황에서의 높은 탐지 성능과 안정성을 향상시키고자 한다.

Ⅱ. 연구내용과 방법

2.1 데이터 수집 및 1D 데이터 전처리

본 연구에서는 Kaggle에서 제공하는 Audio-based Violence Detection Dataset [3]을 사용했다. 해당 데이터 셋은 '폭력'과 '비폭력' 두 가지 레이블로 구성된 영어 음성 데이터를 포함한다. 비폭력 데이터는 1시간 길이의 강연 내용이 담긴 데이터 4개로 구성되어 비폭력을 나타내기에 다양성이 부족하다고 판단하여 일부만 사용하고 유튜브에서 비폭력 영어 데이터를 추가로 수집했다.

영어 데이터뿐만 아니라 한국어 음성 데이터도 포함하기 위해, 유튜브에서 한국어 음성을 수집했으며, 영어와 한국어 음성을 구분 없이 '폭력'과 '비폭력', 두 가지 레이블로 통합한 후 전처리를 진행했다. 먼저, 폭력 데이터의 경우, 폭력과 관련이 없는 내용을 수작업으로 제거하고, noisereduce 라이브러리 [4]를 사용하여 오디오 신호를 주파수 도메인으로 변환한 후 특정 주파수 대역에서 발생하는 노이즈를 제거했다. 또한, pydub 라이브러리 [5]를 사용하여 오디오 무음 처리를 진행했으며 최소 무음 길이 0.5초, 오디오가 무음으로 간주되는 임계값을 -40 dBFS으로 설정하여 학습에 적합한 데이터로 전처리했다

전처리된 데이터는 학습, 검증, 테스트 데이터 각각 7:2:1의 비율로 분할하고, 모든 데이터를 3초로 통일하여 모델 학습에 사용했다.

2.2 2D 데이터 전처리

1D 오디오 데이터를 2D 스펙트로그램 형태로 변환하여 2D CNN 모델의 성능을 극대화하고자 했다. 이 과정에서 인간의 청각 특성이 반영된 Mel Scale로 변환한 멜 스펙트로그램(Mel Spectrogram) [6]을 사용하여, 저주파에서 높은 해상도를 추출했다. 이는 일상 속 발생하는 폭력적인 소리에서 저주파 대역의 특징을 효과적으로 분석하는 데 유리하다. 따라서, 본 연구에서는 폭력 오디오 데이터의 특징으로 멜 스펙트로그램을 사용했다,

2.3 1D, 2D 모델 학습 및 비교

본 연구는 오디오 기반 데이터와 특히, 멜 스펙트로그램과는 ResNet18 모델의 사전학습된 데이터에 특성이 맞지 않다고 판단했다 [7]. 따라서 ResNet18 모델을 기반으로 모든 가중치를 처음부터 학습시켰으며, 이를 통해 오디오 데이터 모델을 구축했다.

1D 오디오 데이터와 2D 스펙트로그램 기반 데이터를 사용한 Residual Learning 기반 CNN 모델의 성능을 비교하기 위해, 두 모델의 파라미터 수를 약 400만 개로 동일하게 설정한 후 학습을 진행했다. 동일한 파라미터 수와 조건에서 학습을 수행해 데이터 형태에 따른 모델 성능 차이를 보다 공정하게 평가하고자 한다.

초기 입력 데이터는 1채널의 오디오 신호로, Conv1d 레이어를 사용하여 64개의 출력 채널로 변환했다. 이후 BatchNorm1d와 MaxPool1d 레이어를 적용하여 데이터의 정규화 및 다운 샘플링을 수행했다. Residual Block을 활용하여 깊은 네트워크에서도 기울기 소실 문제를 완화하였

으며, 각 잔차 블록은 두 개의 Conv1d와 BatchNorm1d로 구성되어, 총 37개의 레이어를 쌓았다. 이후 입력과 출력을 더하는 방식으로 잔차 학습을 진행했다.

2D 학습을 위해 음성신호의 주파수 성분을 시간에 따라 시각적으로 표현하고, 짧은 시간 동안 주파수 변화를 분석하는 STFT(Short Time Fourier Transform)를 사용하여 1D 오디오 데이터를 2D 이미지 형태로 변환했다. 이어서, Conv2d 레이어로 1D 학습 모델과 동일한 구조로 잔차 블록을 쌓고, 모델의 파라미터 수를 맞추기 위해 입력 이미지를 38개의 출력 채널로 변환하여 네트워크를 구성했다.

Ⅲ. 실험 결과

하이퍼 파라미터는 표 1과 같이 동일하게 하고 학습률(Learning Rate)을 변경해 모델 학습을 한 후 비교해 본 결과, 표 2와 같이 학습률 0.001, 0.005에서 2D가 85.31로 1D보다 높은 성능을 보인다.

표 1. 딥러닝 모델의 하이퍼 파라미터 설정

Batch Size	32
Epoch	50
Scheduler	ReduceLROnPlateau
Optimizer	Adam

표 2. ResNet18로 학습한 1d, 2d 성능 비교

Model	Learning Rate	Accuracy
1D	0.001	79.02
2D		**83.92**
1D	0.005	59.44
2D		**85.31**

이후 실험에서는 표 1에서 사용한 하이퍼 파라미터를 조정하고, 데이터 증강 기법을 적용하여 최적의 성능을 찾고자 하였다. 학습률은 0.003으로 설정하였으며, 학습 스케줄러로는 StepLR을 사용했다. 데이터 증강 기법으로 TimeMasking을 사용해 본 결과 time_mask_param이 40일 때, 옵티마이저가 Adam일 때 가장 우수한 성능을 보였다. 이는 Adam이 기울기 업데이트 과정에서 적응형 학습률을 사용하므로 더 빠른 수렴과 안정적인 학습이 가능하기 때문이다.

마지막으로, 이전 실험에서 입력 이미지를 38개 채널로 변환하여 학습한 것을 64개 채널로 변경함으로써 채널 확장을 통해 네트워크가 더 복잡한 특징을 학습할 수 있도록 도와주었다. 학습에서 Accuracy, Loss, Matrix 지표로 성능을 평가했고, 최종적으로 표 3과 같이 테스트 정확도가 90.91%의 성능을 보인다.

표 3. 2D 모델 성능 비교

Scheduler	Augmentation	In_Channels	Accuracy
ReduceLROnPlateau	-	38	83.92
StepLR	-	38	88.11
StepLR	TimeMasking	38	89.51
StepLR	TimeMasking	64	**90.91**

표 4. Confusion Matrix

		Predicted	
		Positive	Negative
Actual	Positive	65	11
	Negative	2	65

Ⅳ. 결론

본 연구에서는 오디오 데이터를 활용한 폭력 탐지 모델을 제안하며, 1D 오디오 신호와 2D 스펙트로그램 변환 데이터를 각각 사용해 Residual Network 기반 CNN 모델의 성능을 비교했다. 영어와 한국어 음성 데이터를 통합하여 학습을 진행했으며, 데이터 증강 기법과 하이퍼 파라미터 최적화를 통해 성능을 개선했다. 또한, ResNet18 모델을 기반으로 모든 가중치를 처음부터 학습시켜 오디오 데이터 모델을 구축했다. 그 결과, 2D 스펙트로그램을 활용한 모델이 1D 모델보다 더 높은 성능을 보였으며, 최종적으로 90.91%의 테스트 정확도를 기록했다.

ACKNOWLEDGMENTS

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

REFERENCES

[1] https://www.index.go.kr/unity/potal/main/EachDtlPageDetail.do?idx_cd=1741

[2] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computervision and pattern recognition (pp.770-778).

[3] https://www.kaggle.com/datasets/fangfangz/audio-based-violence-detection-dataset

[4] https://pypi.org/project/noisereduce/3.0.2/

[5] https://pydub.com

[6] https://www.dbpia.co.kr/pdf/pdfView.do?nodeId=NODE11516239(Mel-Spectrogram과 MFCC를 이용한 딥러닝 기반 딥보이스 탐지시스템 개발에 관한 연구-참고)

[7] 장윤영, 배진희, 임준식. (2019). VGG16 모델을 기반으

로 심음의 Mel Spectrogram을 사용한 심장 질환 분류. 한국소프트웨어종합학술대회. 한국정보과학회 2019 한국 소프트웨어종합학술대회 논문집, 1,537-1,539.