임베디드 환경에서 딥 러닝을 이용한 실시간 병원성 기침 음성 분류

윤정수, 김대원, 김유진, 이현빈*

국립한밭대학교, 한국전자통신연구원, 한국전자통신연구원, *국립한밭대학교

yunjs0126@naver.com, dwk@etri.re.kr, youjin@etri.re.kr, *bean@hanbat.ac.kr

Real-time Pathogenic Cough Sound Classification using Deep Learning in Embedded Environments

Jungsu Yun, Dae-Won Kim, Youjin Kim, Hyunbean Yi*

Hanbat National Univ., Electronics and Telecommunications Research Institute, Electronics and Telecommunications Research Institute, *Hanbat National Univ.

요 약

본 논문은 임베디드 환경에서 실시간으로 병원성 기침 음성을 분류하는 딥 러닝 모델을 제안한다. 입력 데이터는 딥 러닝 모델의 추론 결과로 배경음, 정상인 모사기침, 병원성 기침의 세개 클래스로 분류한다. 임베디드 환경에서 실시간으로 추론하기 위해 MnasNet을 기반으로 모델의 정확도, 추론 응답시간, 목표 응답시간을 최적화한 EfficientNet을 사용하여 학습 및 평가를 실시하였다. 학습 및 평가 결과 EfficientNet-B3를 사용하였을 때 가장 높은 정확도와 F1-Score를 달성하였으며, 임베디드 환경에서 1초의 데이터를 입력으로 사용하는 경우 실시간으로 병원성 기침 분류가 가능하다.

Ⅰ. 서 론

감기와 독감 같은 호흡기 질환은 매년 유행하며, 특히 코로나-19는 전 세계적인 대유행을 일으켰다. 호흡기 질환은 바이러스, 세균, 진균 등의 다양한 병원균에 의해 발생하며 기침으로 인한 비말 전파로 확산 된다[1]. 기침은 호흡기 질환의 초기 단계에서 발생한다는 것을 고려할 때 병원성 기침을 빠르게 인식하는 것이 중요하다. 최근에는 기침 소리를 이용한 병원성 기침을 분류하는 연구가 활발히 수행되고 있으며[2] 딥 러닝을 이용하여 호흡기 질환을 검출하는 연구도 진행되고 있다[3].

기존의 기침 음성 분류 연구는 이진 분류 문제(Binary classification)로 접근하여 음성 신호에서 기침 소리가 포함되어 있는지 판단하거나, 기침이 포함된 음성 신호에서 병원성 기침 여부를 판단한다[4]. 하지만 실제 환경에서는 배경음과 같은 기타 음성 신호가 포함되어 있어 다중 클래스 분류 문제(Multi class classification)로 접근해야 한다. 또한, 기존의 기침 음성 분류 모델은 고성능 GPU를 사용하여[5] 컴퓨팅 성능이 제한되는 임베디드 환경에서는 실시간 추론이 어려울 수 있다.

따라서 본 논문에서는 임베디드 환경에서 실시간으로 딥 러닝을 이용해 병원성 기침, 정상인의 모사 기침, 배경음을 분류하는 모델을 제안한다. 제안하는 모델은 일정 간격의 기침 음성 신호로부터 특징을 추출한다. 추출된 음성 특징을 딥 러닝 모델의 입력으로 사용하여 세 개의 클래스로 분류한다.

본 논문의 구성은 다음과 같다. 2장에서는 제안하는 딥 러닝 모델에 사용된 음성 신호의 수집 및 전처리 프로세스를 소개한다. 3장에서는 추출된 특징 정보를 사용하여 딥 러닝 모델 학습 결과 및 임베디드 환경에서의 추론 속도를 서술하고 4장에서 결론을 맺는다.

Ⅱ. 데이터 수집 및 전처리

본 연구를 위해 사용된 데이터 세트의 구성은 표 1과 같다. 배경음 및 정상인 모사 기침은 한국인 기침 음성 데이터 세트를 사용하였으며, 병원성 기침은 대전광역시 충남대학교 병원의 호흡기 내과에 방문한 환자의 기침 음성을 수집하여 활용하였다. 수집된 데이터는 1초 내외의 데이터로 구성하여 배경음, 모사 기침, 병원성 기침으로 분류하였다.

표 1. 클래스별 수집 데이터 수

Class	병원성 기침	모사 기침	배경음	Total
Number	422	988	421	1,831

데이터 세트는 딥 러닝 모델의 입력으로 사용하기 위해 전처리를 실시한다. 음성 신호의 주파수는 입력되는 마이크의 성능에 따라 달라짐으로 하나의 샘플레이트(Samplerate)로 일치시켜야 한다. 본 연구에서는 입력 샘플레이트를 44.1 kHz로 결정하고 입력 데이터의 샘플레이트가 44.1 kHz가 아닌 경우 이를 맞춰주는 리샘플링(Resampling) 단계를 거친다. Resampling이 완료된 후 음성 데이터를 모노(Mono) 채널로 변환한 뒤, 음성 신호의 길이를 학습 시간 간격과 일치시킨다. 만약 음성 신호가 학습할 시간 간격보다 짧다면 0으로 패딩(Padding)을 하며, 반대로 음성 신호가 학습 시간 간격보다 길다면 앞에서부터 잘라 1초 길이만 사용한다.

수집된 데이터 세트 수는 모델별 입력과 출력 사이의 관계를 학습하기에 충분하지 않아 과대 적합(Overfitting)이 나타날 수 있다. 본 연구에서는 원본 오디오에 배경 잡음을 추가하는 어 데이터 증강(Data Augmentation) 기법을 사용하였다. 배경 잡음은 에어컨, 공항, 병원 소리 데이터 세트를 사용한다[6]. 음성 신호에 배경 잡음을 추가하기 위해 신호 대비 잡음 비(Signal-to-noise ratio, SNR)를 계산하여 음성 신호를 결합한다. 배경 잡음이 추가된 이후 특징을 추출한다. 음성 신호의 특징정보는 Mel-Spectrogram, MFCC(Mel-Frequency Cepstral Coefficient), CQT(Constant-Q Transform)를 각각 추출한 뒤 결합하여 하나의 채널로 사용한다.

Ⅲ. 모델 학습 및 평가

3.1 모델 선정 및 학습

최신의 CNN 모델들은 높은 정확도를 달성할 수 있지만 모델의 크기가 커짐에 따라 추론 속도가 감소하고 연산 비용이 증가하고 있다. 제한된 자원에서 실시간으로 추론이 이루어지기 위해서는 정확도와 추론 시간 사이에서

타협하여야 한다[7]. 본 연구에서는 MnasNet을 기반으로 모델의 정확도, 추론 응답시간, 목표 응답시간을 최적화한 EfficientNet을 사용하였다[8]. EfficientNet은 딥 러닝 모델의 깊이, 너비, 입력 데이터의 크기 사이의 관계를 compound coefficient를 사용해 최적의 모델을 구성하고 높은 성능을 달성할 수 있도록 구현된 모델이다.

손실함수는 Cross Entropy Loss를 사용하였으며, Cosine Annealing 학습률 스케줄러를 사용하였다. NVIDIA RTX A5000 GPU를 사용하여 100 epoch의 학습 결과 정확도와 Loss 곡선은 그림 1과 같다.

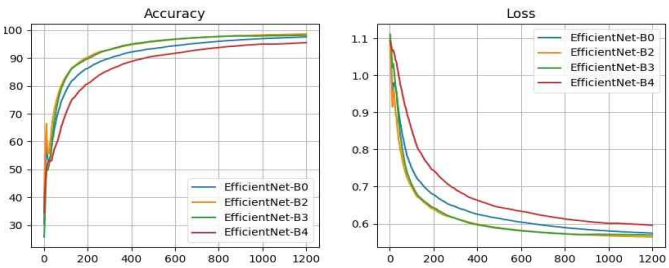

그림 1. 모델 학습 곡선

3.2 모델 성능 평가

모델의 성능을 평가하기 위해 배경음, 정상인 모사 기침, 병원성 기침의 음성 신호를 각각 50개씩 총 150개의 데이터를 사용하였고, 평가 결과는 표 2와 같다. 모델의 크기가 커짐에 따라 추론 정확도와 F1-Score는 향상되는 것을 확인하였고 EfficientNet-B3 모델을 사용하였을 때 가장 높은 정확도와 F1-Score를 달성할 수 있음을 확인할 수 있다. EfficientNet-B4 모델을 사용할 때 정확도가 급격하게 낮아지는 것을 알 수 있다. 그림 2는 배경음, 정상인 모사 기침, 병원성 기침의 MFCC 특징을 추출한 결과이다. 정상인 모사 기침과 병원성 기침은 특정 대역폭에서 유사한 특징이 나타나 데이터의 복잡성이 낮아지는 현상으로 과대 적합이 나타난 것으로 판단된다.

표 2. 모델 평가 결과

Model	Acc.	Loss	Precision	Recall	F1-Score
EfficientNet-B0	98.000	0.572	0.980	0.980	0.980
EfficientNet-B2	95.333	0.601	0.959	0.953	0.954
EfficientNet-B3	**98.667**	**0.565**	**0.987**	**0.987**	**0.987**
EfficientNet-B4	91.333	0.637	0.927	0.913	0.914

배경음 정상인 모사 기침 병원성 기침

그림 2. 클래스별 MFCC 추출 결과

3.3 임베디드 환경 추론 속도 분석

임베디드 환경에서 실시간으로 모델 추론이 가능한지 확인하기 위해 3.2에서 가장 높은 성능을 보인 EfficientNet-B3 모델을 RTX A5000, i7-10700, NVIDIA Jetson Orin, NVIDIA Jetson Nano, Raspberry Pi 4b를 사용하였으며, 평균 추론 시간 측정 결과는 표 3과 같다. 모델의 입력에 사용된 음성 신호는 1 초로 구성되어 있어, Raspberry Pi 4b와 같은 저전력 임베디드 환경에서도 평균 0.8 초 이내에 추론이 이루어져 실시간 추론이 가능함을 확인할 수 있다.

표 3. 임베디드 환경 추론 시간

단위: 초

Content	Min	Max	Average
RTX A5000	0.013	0.729	0.018
i7-10700	0.125	0.138	0.129
Jetson Orin	0.032	0.196	0.033
Jetson Nano	0.100	0.930	0.100
Raspberry Pi	0.789	0.811	0.795

Ⅳ. 결론

본 논문에서는 임베디드 환경에서 실시간으로 병원성 기침 음성을 분류하는 모델을 제안하였다. 다양한 환경에서 수집된 배경음, 기침 음성 데이터를 사용하여 음성 특징을 추출하고 이를 임베디드 환경에서 동작할 수 있는 딥 러닝 모델을 사용하여 분류 정확도를 향상하였다. 향후에는 속 기침과 겉 기침 분류 학습과 이미지 변환 없이 시계열 분석 기반의 인공지능 모델을 구현하여 정확도와 속도를 향상시킬 예정이다.

ACKNOWLEDGMENT

이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2021-0-00725,밀폐공간내 감염병 위험도 감시를 위한 멀티모달 센싱 기반 감시지능 시스 템 기술 개발)

참 고 문 헌

[1] S. E. Park, “Epidemiology, Virology, and Clinical Features of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2; Coronavirus Disease-19),” *Pediatric Infection and Vaccine*, vol. 27 no.1, pp. 1-10, Mar. 2020.

[2] K. S. Alqudaihi, N. Aslam, I. U. Khan, A. M. Almuhaideb, S. J. Alsunaidi, N. M. Abdel Rahman Ibrahim, F. A. Alhaidari, F. S. Shaikh, Y. M. Alsenbel, D. M. Alalharith, H. M. Alharthi, W. M. Alghamdi, and M. S. Alshahrani, “Cough Sound Detection and Diagnosis Using Artificial Intelligence Techniques: Challenges and Opportunities,” *IEEE Access*, vol. 9, pp. 102327-102344, Jul. 2021.

[3] S. J. Yoo, and J. Y. Kim, “Detection of COVID-19 from cough patterns,” *KCIS Fall Conference 2023*, pp, 1299-1300, Pyeongchang, Korea, Feb. 2022.

[4] M. K. Kim, G. W. Kim, K, and H. Choi, “A COVID-19 Diagnosis Model based on Various Transformations of Cough Sounds,” *Journal of Intelligence and Information Systems*, vol. 29, no. 3, pp. 57-78, Sep. 2023.

[5] K. Habashy, J. Valdés, M. Cohen-McFarlane, P. Xi, B. Wallace, R. Wallace, R. Goubran, and F. Knoefel, "Cough Classification Using Audio Spectrogram Transformer," *2022 IEEE Sensors Applications Symposium (SAS)*, Sundsvall, Sweden, Aug. 2022.

[6] C. K. A. Reddy, E. Beyrami, J. Pool, R. Cutler, S. Srinivasan, and J. Gehrke, “A scalable noisy speech dataset and online subjective test framework,” *ISCA Interspeech 2019*, pp. 1816-1820, Sep. 2019.

[7] M. Tan, et al., "MnasNet: Platform-Aware Neural Architecture Search for Mobile", 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, pp. 2815-2823, Jun, 2019.

[8] M. Tan, Q, and V. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, ” *arXiv*, May. 2019.