한국음향학회지 제42권 제3호 pp. 221~225 (2023)
The Journal of the Acoustical Society of Korea Vol.42, No.3 (2023)
https://doi.org/10.7776/ASK.2023.42.3.221

pISSN : 1225-4428
eISSN : 2287-3775

트랜스포머를 이용한 음성기반 코비드19 진단

Audio-based COVID-19 diagnosis using separable transformer

강승태,[1] 장길진[1†]

(Seungtae Kang[1] and Gil-Jin Jang[1†])

[1]경북대학교 전자전기공학부

(Received April 18, 2023; accepted May 17, 2023)

초 록: 본 연구에서는 코로나 바이러스 감염증은 음성만으로 빠르게 진단하는 효율적인 방법을 제안하였다. 기존의 딥러닝 기반 방법들의 연산시간과 대용량 학습자료 요구조건을 완화하기 위해서 Separable Transformer(SepTr)의 구조를 개선하여 파라미터의 수를 대폭 감소시키고 빠른 진단을 가능하게 하는 새로운 Strided Convolution Separable Transformer(SC-SepTr)를 제안하였다. 공개 음향 데이터인 Coswara에 대하여 실험을 수행한 결과 제안된 방법은 상대적으로 소규모의 학습자료에 대해서도 Area Under the Curve(AUC) 성능을 보장하면서도 신속하게 진단을 수행할 수 있음을 보였다.

핵심용어: 코로나바이러스감염증, 기침소리, 숨소리, 트랜스포머, 분리형트랜스포머

ABSTRACT: In this paper, we proposed an efficient method for rapid diagnosis of COVID-19 by voice. A novel Strided Convolution Separable Transformer (SC-SepTr) is proposed by modifying the conventional Separable Transformer (SepTr) for audio signal recognition. The proposed method reduces the memory and computational requirements to enable rapid diagnosis of COVID-19. As a result of experiments on Coswara, it was shown that the proposed method perform rapid diagnosis with guaranteeing Area Under the Curve (AUC) performance even for a relatively small amount of learning data.

Keywords: COVID-19, Cough, Breathing, Transformer, Separable transformer

PACS numbers: 43.80.Qf, 43.70.Dn

I. 서 론

세계 보건 기구의 통계에 따르면 코로나 바이러스 감염증 Coronavirus disease(COVID)-19은 2019년 하반기에 발생하여 전 세계 인구의 10 % 이상을 감염시키며 전 세계적인 유행병으로 발전하였다.[1] COVID-19의 전파를 막기 위한 선제적 검사는 유전자 기반 중합 효소 연쇄 반응(Polymerase Chain Reaction, PCR) 검사와 신속항원검사(Rapid Antigen Test, RAT; Antigen Rapid Test, ART)가 있다.[1] 하지만, 표본채취 후 결과를 받아보는 데까지 시간이 오래 걸리고, 특정 검사소에 가야 하는 등 신속한 검사가 이루어지기 어렵다. 이를 보완하기 위해 대상자의 기침소리로 COVID-19를 감별하기 위한 다양한 데이터셋이 구축되었으며,[2-4] 이를 이용한 진단에 관해 연구가 진행되고 있다.[5,6]

딥러닝을 이용한 음향신호 인식은 트랜스포머를 음향신호에 적합하게 구현한 오디오 트랜스포머(Audio Spectrogram Transformer, AST)[7]와 이를 유사한 모델들이 사용되고 있다. 본 논문에서는 AST를 개선시킨 Separable Transformer(SepTr)[8]을 이용하여 대상자의 음성으로 COVID-19을 진단하는 방법을 제안한다. 본 연구에서는 기존의 SepTr 모델 구조를 변경하여 연산량과 메모리 사용량을 획기적으로 줄이는 구조를 제안하고, 이를 음성기반 COVID-19 진

†**Corresponding author:** Gil-Jin Jang (gjang@knu.ac.kr)
School of Electronic and Electrical Engineering, Kyungpook National University. 80 Daehakro, Daegu 41566, Republic of Korea
(Tel: 82-53-950-5517, Fax: 82-53-950-5501)

단에 적용하는 방법을 제안하였다. 제안된 방법의 유효성 검증을 위해 Coswara('coronavirus'와 소리를 의미하는 산스트리트어 'swara'의 조합)[3] 데이터베이스로 실험이 시행되었으며 기존 AST 대비 대부분 제안된 방법이 높은 진단 성능을 보였다.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 음성기반 COVID-19 진단방법들을 소개한다. 3장은 제안된 방법에 대하여 자세히 설명한다. 4장에서 실험환경 및 결과를 열거하고 5장에서 결론을 맺는다.

II. 관련연구

COVID-19가 빠르게 확산되면서 음향신호만으로 감염을 신속하게 진단하려는 시도가 여러 차례 있었다. 진단 성능을 평가하기 위해서 COUGHVID(기침의 영단어 'cough'와 'COVID'의 조합)[2] 및 Coswara[3] 등 연구에 사용될 수 있는 여러 가지 음성 데이터베이스가 공개되었으며 이를 기반으로 아양한 음성 기반 COVID-19 진단방법들이 제안되었다.[3,9] 하지만 학습 데이터의 양이 충분하지 않아 이를 극복하기 위해 비교사학습 또는 자기지도 학습을 적용하는 방법도 연구되고 있다.[10,11]

컴퓨터 비전 분야에서는 비전 트랜스포머(Vision Transformer, ViT)가 제안되었다.[12] 오디오 분류에서는 ViT와 유사한 구조를 구축하고 영상 대신 음향 스펙트로그램의 패치 순열을 입력으로 사용하는 AST가 제안되었다.[7] AST는 입력 스펙트럼을 16×16 크기의 2차원 패치들로 나누고 패치들의 순열을 원래 순서에 따라 트랜스포머의 입력으로 사용한다. 이 방식은 입력의 크기가 커질수록 순열의 길이가 길어져 연산량과 메모리 사용량이 증가하는 문제가 있고, 신속한 COVID-19 진단에 적용되기 어렵다.

III. 제안된 방법

AST는 매우 신뢰성 있는 결과를 제공하지만 일반적인 신경회로망과 마찬가지로 주어지는 학습자료의 양에 따라 성능의 편차가 발생된다. 자연어 처리, 이미지 처리 영역에서는 이를 해결하기 위해 대량의 데이터셋을 이용한 사전학습 방식을 사용한다. 하지만 의료적인 오디오 분류 작업에 사용되는 오디오의 특성상 사전학습을 할 수 있을 정도의 대량의 데이터셋을 같은 도메인에서 구하기는 어렵다는 문제가 있다. 따라서 구조를 최적화하여 적은 데이터셋에서도 성능을 보장하는 모델 구조가 필요하다.

매개 변수의 수와 메모리 사용량을 줄이기 위해 분리 가능한 트랜스포머가 제안되었다.[8] SepTr은 주파수 축과 시간 축의 상관관계를 분리하여 별도의 신경망 층에서 처리하며, 시간축과 주파수축을 나누어 임베딩하기 때문에 서로 상이한 특성을 가지는 시간과 주파수 축의 연속성을 별개로 모델링할 수 있고, 음성신호에 보다 적합하다. 다만 패치의 크기를 작게 설정하여 세밀한 부분을 볼 수 있지만, 필요한 연산량과 메모리가 크게 증가하게 되었다.

Fig. 1은 기존의 AST와 SepTr을 제안된 Strided Convolution Separable Transformer(SC-SepTr) 구조와 비교하였다. AST는 시간과 주파수 영역을 하나의 벡터로 구성하고, 이를 트랜스포머의 입력으로 사용하였다. 하지만 AST는 시간 및 주파수 성분에 정확하게 주의값을 주기 어렵다. 이를 개선한 SepTr은 시간과 주파수 축을 나누고, 각각에 서로 다른 시간축 트랜스포머와 주파수축 트랜스포머를 적용하였다. 반복되는 구조에서는 입력의 크기를 줄이는 평균축소과 원래의 크기로 확장이 요구된다. SepTr은 AST에 비하여 작은 패치크기를 사용하는 것과 원래 크기로 확장하는 부분에 의해 모델에 필요한 리소스가 증가하게 된다. 본 연구에서는 기존의 SepTr의 구조를 개선하여 적은 리소스를 사용하면서도 신뢰성 있는 결과를 얻을 수 있도록 하였다. Fig. 1-(c)와 같이 반복되는 구조 이전에 $p \times p$ 스트라이드 합성곱(strided convolution)을 추가하고 반복구조 내의 선형투영과 평균축소는 삭제하였다. 이에 따라 반복구조 마지막에 확장도 필요가 없어지기 때문에 각 블록을 처리한 후 다시 업샘플링하는 부분이 제거된다. 이에 따라 반복구조가 매우 간단해지고 학습시간과 메모리 사용량을 크게 줄일 수 있다. 기존의 평균축소는 입력 패치의 크기에 비례하여 정보의 손실이 증가되지만, 스트라이드 합성곱 구조에서는 이러한 문제점을 최소화 할 수 있다.

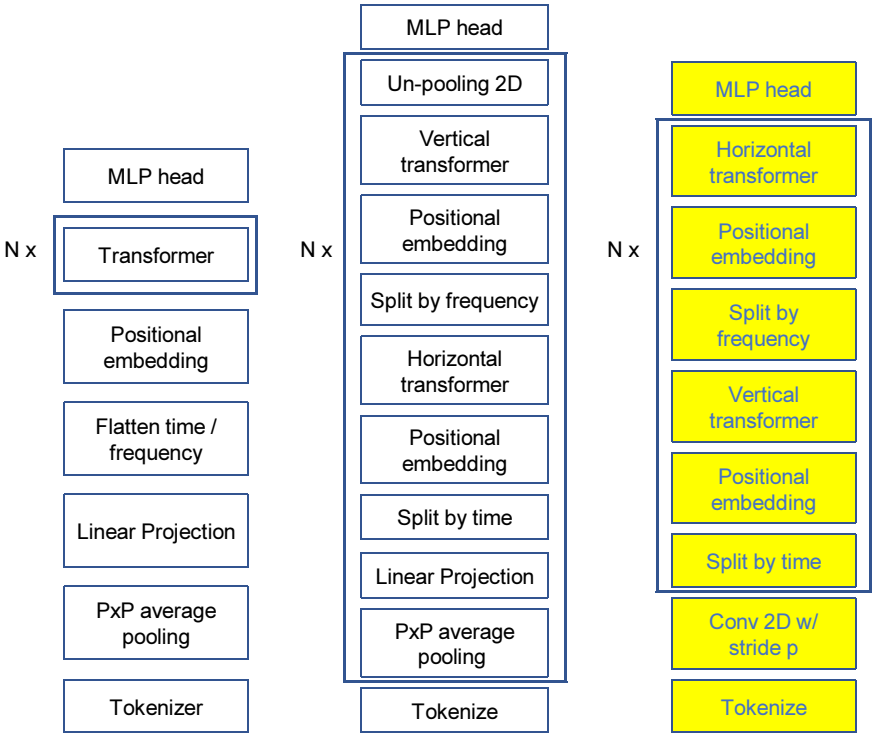

Fig. 1. (Color available online) Comparison of AST and the conventional SepTr, and the proposed SC-SepTr architectures.

IV. 실험 및 결과

본 연구에서는 제안된 방법과 기존의 방법들을 비교 평가하기 위하여 인도과학원의 Coswara[3] 데이터베이스를 사용하였다. 포함된 비음성 자료는 기침, 호흡, 숫자 세기, 모음 연장 발화를 정상인과 COVID-19 감염자들로부터 비대면으로 수집한 녹취샘플들이며 발화자의 나이, 성별, 기저질환 등의 정보도 제공하고 있다. 학습자료와 평가자료가 따로 구분되어 있지 않기 때문에 총 현재 2,747개의 발화 샘플들을 임의로 학습, 검증, 평가 자료의 비율을 6:2:2 로 나누어 실험을 진행하였다.

우리는 이 데이터셋을 16 kHz로 샘플링 주파수로 통일하여 실험에 사용하였다. 실험은 각 입력 방식에 대해 같이 혹은 따로 시행하였으며, 0.5 s 단위로 임의로 추출된 발화음성에대해 COVID-19 환자판단 여부 정확도로 성능을 평가하였다. 입력 특성은 128 차원의 멜단위 필터뱅크 에너지가 구현된 신경망의 입력으로 주어졌다. 기존의 AST와 오디오 음성에 적절하게 구성된 제안된 SepTr 모델에 대하여 비교 성능평가가 시행되었다. AST는 SepTr, SC-SepTr과 비교를 위해 선행학습 가중치를 사용하지 않고 처음부터 학습한 모델로 성능을 평가하였다.

Table 1. Comparison of model sizes of AST, SepTr, and SC-SepTr.

	AST	$SepTr_{org}$ [8]	SepTr	SC-SepTr
Computation time per sample (ms)	16.6	-	9.9	7.8
Required memory (MB)	14069	OOM	10863	4171
Number of parameters (M)	87.3	40.3	39.4	39.5

Table 1에서 실험에 사용한 각 모델의 크기를 비교하였다. 처리 시간과 메모리 사용량은 NVIDIA RTX 3090에서 실행하여 측정하였다. 기존의 패치크기 8 × 8로 구현한 $SepTr_{org}$[8]는 중간과정 연산에 사용되는 특징의 크기가 커서 NVIDIA RTX 3090의 24 GB 메모리에서 실행이 불가능하였다(Out Of Memory, OOM). 따라서 본 실험에서는 AST와 같이 패치크기를 16으로 키운 SepTr을 사용하였다. 또한 SepTr, SC-SepTr과

Table 2. Comparision of AUC results of AST, SepTr, and SC-SepTr.

Sound type	AST	SepTr	SC-SepTr
Counting-fast	0.66	0.67	0.70
Counting-normal	0.62	0.63	0.72
Cough-heavy	0.70	0.62	0.68
Cough-shallow	0.66	0.61	0.68
Breathing-deep	0.62	0.63	0.69
Breathing-shallow	0.69	0.62	0.69
Vowel-a	0.62	0.62	0.68
Vowel-o	0.62	0.59	0.63
Vowel-e	0.62	0.61	0.64
Average	0.64	0.62	0.67

AST 모두 6개의 블록을 사용하였고 패치크기는 모두 같은 16 × 16을 적용하였다.

AST는 같은 블록 수 기준 가장 많은 파라미터 수와 큰 연산시간, 메모리 사용량이 측정되었다. SepTr과 SC-SepTr은 주파수 축과 시간 축으로 나누어 학습하는 구조를 통해 파라미터수를 절반 정도만을 요구한다. 또한 SC-SepTr이 SepTr보다 적은 연산 시간과 메모리를 요구한다. 그 이유는 반복되는 풀링과 언풀링을 제거하였기 때문이다.

정상인과 감염자의 숫자의 차이 문제를 해결하기 위하여 무작위 추출이 적용되었다. 하이퍼파라미터들은 검증 자료의 성능으로 선택되었다. 결과는 Table 2과 같다. 성능비교는 이진분류기의 성능평가에 널리 사용되는 Area Under the Curve(AUC)를 이용하였다.[13] AUC는 거짓 긍정률(False Positive Rate, FPR)을 변경하면서 구한 참 긍정률(True Positive Rate, TPR)을 적분한 면적으로 정의되며, 이진 분류기의 문턱치 설정에 영향받지 않는 성능을 평가하는데 매우 유용하다. Table 1에서 AST가 SC-SepTr의 약 2배의 파라미터를 가지고 있음에도 불구하고 cough-heavy와 breathing-shallow 두 가지 입력을 제외하면 SC-SepTr을 이용하는 것이 더 높은 AUC를 기록하였다. 이는 학습 데이터의 수에 비해 모델의 크기가 너무 커 과적합이 발생한 결과로 보인다. SepTr보다 SC-SepTr이 적은 수의 파라미터 증가만으로 모든 입력 형식에 대하여 좋은 AUC을 보여주었다. 이는 패치 사이즈가 큰 경우 SepTr의 평균축소에 비해 SC-SepTr의 스트라이드 합성곱이 더 효과적임을 보여 준다.

Table 3. Comparison of AUC results of AST and SC-SepTr with and without pretraining.

	Without pretraining		With pretraining	
	AST	SC-SepTr	AST	SC-SepTr
Cough-heavy	0.70	0.68	0.58	0.58
Cough-shallow	0.66	0.68	0.67	0.67

Coswara는 2,747개의 발화만으로 구성되어 있으며, 본 실험에서는 이를 6:2:2로 학습, 검증, 테스트 데이터로 나누어 사용했기 때문에 학습자료의 양이 상대적으로 부족하다. 따라서, Table 1의 결과가 적절한지 검증하기 위하여 추가적인 사전학습을 실시하는 실험을 수행하였다. 일반적인 음성이나 음악은 Coswara의 음원과 종류가 매우 다르기 때문에 본 연구에서는 기침소리 기반 COVID-19 분류를 위한 COUGHVID으로 사전 학습한 가중치를 이용하여 추가 실험을 진행하였다. COUGHVID 데이터셋은 정상인과 COVID-19 환자의 약 25,000개의 기침 소리를 크라우드 소싱으로 수집한 데이터셋이으로 다른 사전학습을 이용한 COVID-19 판별기에서 주로 사용되는 데이터셋이다. 우리는 이 데이터셋을 16 kHz로 샘플링 주파수를 변환하여 AST와 SC-SepTr을 사전학습하고 성능을 평가하였다. 결과는 Table 3에서 AUC로 비교하였다. cough-heavy의 경우는 오히려 사전학습 이후 AUC 0.12, 0.1의 성능저하가 보였다. cough-shallow는 AST는 AUC 0.01 증가, SC-SepTr는 0.01 저하되었다. Coswara의 cough-heavy는 일부러 강하게 기침을 하게 하여 녹음하였기 때문에 COUGHVID의 음원과 달라 성능이 오히려 저하되었다. Coswara cough-shallow는 COUGHVID와 음원특성이 비슷하여 성능이 저하되진 않았지만 사전학습에 의한 성능 향상은 없었다. 따라서 Coswara만으로도 학습해도 충분한 성능을 얻을 수 있다.

V. 결 론

본 연구에서는 코로나 바이러스 감염증을 음성으로 빠르게 진단하는 효율적인 방법을 제안하였으며 기존의 AST와 SepTr을 개선하여 적은 학습데이터 환경에서 효과적인 COVID-19 감염병 진단방법을 제안하였다. 제안된 SC-SepTr은 시간-주파수 영역으로

분리하는 SepTr의 특징은 유지하면서 신속한 진단이 가능하게 한다. 공개 감염병 데이터인 Coswara에 대하여 실험을 수행한 결과 제안된 방법은 기존의 AST 대비 성능향상을 보임으로써 제안된 방법의 유효성을 검증하였다.

감사의 글

본 연구는 행정안전부의 재원으로 방역연계범부처감염병연구개발 사업단의 지원을 받아 수행되었습니다(과제고유번호 : 20016180, 100 %).

References

1. *World Health Organization Official COVID-19 info Official Website*, https://www.who.int/covid-19, (Last viewed May 23, 2023).
2. L. Orlandic, T. Teijeiro, and D. Atienza, "The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms," Sci. Data, **8**, 1-10 (2021).
3. N. K. Sharma, P. Krishnan, R. Kumar, S. Ramoji, S. R. Chetupalli, N. R., P. K. Ghosh, and S. Ganapathy, "Coswara - a database of breathing, cough, and voice sounds for COVID-19 diagnosis," Proc. Interspeech, 4811-4815 (2020).
4. D. T. Pizzo and S. Esteban, "IATos: AI-powered pre-screening tool for COVID-19 from cough audio samples," arXiv:2104.13247 (2021).
5. A. Mallol-Ragolta, H. Cuesta, E. Gomez, and B. Schuller, "Multi-type outer product-based fusion of respiratory sounds for detecting COVID-19," Proc. Interspeech, 2163-2167 (2022).
6. V. S. Nallanthighal, A. Harma, and H. Strik. "COVID-19 detection based on respiratory sensing from speech," Proc. Interspeech, 2498-2502 (2022).
7. Y. Gong, Y.-A. Chung, and J. Glass, "AST: audio spectrogram transformer," Proc. Interspeech, 571-575 (2021).
8. N. C. Ristea, R. T. Ionescu, and F. S. Khan, "SepTr: separable transformer for audio spectrogram processing," Proc. Interspeech, 4103-4107 (2022).
9. N. K. Sharma, S. R. Chetupalli, D. Bhattacharya, D. Dutta, P. Mote, and S. Ganapathy, "The second DiCOVA challenge: Dataset and performance analysis for diagnosis of COVID-19 using acoustics," Proc. ICASSP, 556-560 (2022).
10. X.-Y. Chen, Q.-S. Zhu, J. Zhang, and L.-R. Dai, "Supervised and self-supervised pretraining based covid-19 detection using acoustic breathing /cough/ speech signals," Proc. ICASSP, 561-565 (2022).
11. T. Dang, T. Quinnell, and C. Mascolo, "Exploring semi-supervised learning for audio-based COVID-19 detection using FixMatch," Proc. Interspeech, 2468-2472 (2022).
12. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," Proc. ICLR, 1-22 (2021).
13. K. Lee and C. H. Lee, "Abnormal signal detection based on parallel autoencoders" (in Korean), J. Acoust. Soc. Kr. **40**, 337-346 (2021).

저자 약력

▸ 강 승 태 (Seungtae Kang)

2018년 2월 : 경북대학교 전자공학부 학사
2020년 2월 : 경북대학교 전자공학부 석사
2020년 3월 ~ 현재 : 경북대학교 전자전기공학부 박사과정

▸ 장 길 진 (Gil-Jin Jang)

1997년 2월 : KAIST 전산학과 학사
1999년 2월 : KAIST 전산학과 석사
2004년 2월 : KAIST 전산학과 박사
2006년 8월 : 삼성종합기술원 전문연구원
2009년 10월 : 박사후연구원, Univ. California, San Diego, USA
2014년 2월 : 울산과학기술대학교(UNIST) 전기전자컴퓨터공학부 조교수
2014년 3월 ~ 현재 : 경북대학교 전자전기공학부 조교수/부교수