v. 2, n. 33, 2022

International Journal of Health Science

DETECTION OF COVID-19 IN RESPIRATORY SOUNDS USING END-TO-END DEEP AUDIO EMBEDDINGS

Carlos A. Galindo-Meza
DESI Instituto Tecnologico de Estudios Superiores de OccidenteTlaquepaque, Jal., Mexico

Juan A. del Hoyo Ontiveros
Electrónica e Informática
Universidad Autónoma de Guadalajara
Zapopan, Jal., Mexico

Jose I. Torres Ortega
Programa de IOT
Universidad Alinnco
Santiago de Querétaro, Qro., Mexico

Paulo Lopez-Meyer
Intel Labs
Intel Corporation
Zapopan, Jal., Mexico

Abstract: Due to the COVID-19 worldwide pandemic situation, automatic audio classification research has been of interest for analysis of respiratory sounds. Several deep learning approaches have shown promising performance for distinguishing COVID-19 in respiratory cycles. In this work we explored the usage of transfer learning from a pre-trained end-to-end deep-learning based audio embeddings generator named AemResNet, applied to the classification of respiration and coughing sounds into healthy or COVID-19. We experimented with the publicly available large-scale Cambridge Crowdsourced dataset of respiratory sounds collected to aid diagnosis of COVID-19. Our presented work focuses into 3 experimental tasks: 1) detection of COVID-19 from a combination of breath and cough sounds, 2) detection of COVID-19 from breath sounds only, and 3) detection of COVID-19 from cough sounds only. The experimental results obtained over this respiratory dataset show that a pre-trained audio embedding generator achieves competitive performance compared to the recent published state-of-the-art.
Keywords: Audio classification, cough sounds, COVID-19 detection, deep learning, respiratory sounds, transfer learning.

INTRODUCTION

Coronavirus (COVID-19) is an infectious disease caused by the severe acute respiratory syndrome coronavirus (SARS-CoV-2) virus [1] first detected in Wuhan, China in 2019. On March 2020th, COVID-19 was declared a pandemic by the World Health Organization (WHO). Most people experience moderate respiratory symptoms such as: coughing, fever, and shortness of breath. The first time this novel virus was detected was within a cluster of patients with pneumonia of unknown cause. According to the WHO, 15% of overall COVID-19 patients present a severe pneumonia [1], which is auscultated by a physician listening respiratory sounds through breath and cough. The main purpose of recording respiratory sounds is to find a weakness of hypoventilation which can lead to diagnose the patient illness.

Nowadays, there are several methods proposed to distinguish the respiratory cycles, e.g., identifying a shortness of breath mostly related to pneumonia. The implementation of the most recent approaches on respiratory sound classification includes a recurrent neural network used for lung sound classification in [2] to predict respiratory anomalies is proposed in [3], a deep learning architecture to detect possible lung disease in presented in [4] by classifying respiratory anomalies. A VGG16 CNN for automatic classification of respiratory sounds was proposed in [5] also by means of deep learning.

As well, COVID-19 aimed works have taken part on the research community. The work reported in [6] shows the efforts on the creation of an Android application aimed to collect different sounds from patients such as breath, cough, and speech; with this, they have created a dataset containing more than 459 samples from 378 patients through a crowdsourced methodology, named Cambridge Crowdsourced dataset. In this work, some machine learning (ML) techniques such as Support Vector Machines (SVM) were used as the classifier for COVID-19 detection. In [7], the composition of residual network blocks is used to classify COVID-19 based on audio spectrograms and motivates to a comprehensive follow-up research. On [8], respiratory audio recordings are treated as a visual representation through two different spectrogram configurations and as raw audio, each of these samples are inputted into a CNN layer and the output is concatenated and ensembled to classify

COVID-19. Overall, it can be observed how deep learning is currently leading the state-of-the-art (SOTA) when it comes to audio classification for COVID-19.

In this work we propose the use of an end-to-end (e2e) deep learning-based model to identify healthy breath and/or coughing sounds from COVID-19 ones. We have arranged our work as follows: Section II describes the methodology followed for the implementation of the deep learning audio classification of healthy vs COVID-19 sounds; Section III presents a clear explanation on the experimental setup; Section IV presents the experimental results obtained and the discussion around them; and finally, Section V presents the conclusions drawn from this work.

METHODOLOGY

A deep learning approach for detection of COVID-19 respiratory sounds presented in this work, based on an end-to-end (e2e) convolutional neural network (CNN); this means that no additional audio spectral representation is needed since the time-domain signals are the input to the neural network architecture. This approach seems optimal when considering the dedicated hardware limitations for inference deployment. The core of this work is an ongoing effort of the e2e audio embeddings generator described in previous published works [9]–[12], where pre-trained models are created through an available large audio dataset, that efficiently generate robust audio embeddings aimed for different audio scene and events classification. The proposed e2e CNN architecture is named AemResNet, and it comprises three main blocks as seen in Fig 1: the low-level feature block (LLF) that acts as a front-end learnable feature extraction module, the high-level feature block (HLF) that is trained to become a deep learning-based audio embeddings generator, and a final classification block that is trained with the audio embeddings output by the HLF.

The purpose of the LLF block is to discriminate and extract features based purely from raw audio; this block replaces the visual representation of audio through spectrograms commonly used in most audio classification tasks. In Fig 1, the details of this block are described, where we find two 1-dimensional (1D) strided convolutional layers (Conv), each followed by a batch normalization layer (BN) and a ReLu activation function. The 16 kHz time-domain audio waveform inputted to the LLF block is converted to 128 channels using a time window resolution of 10ms after an added max-pool layer. For each second of audio input, the LLF block creates a [128, 1, 100] dimension tensor which act as a trainable

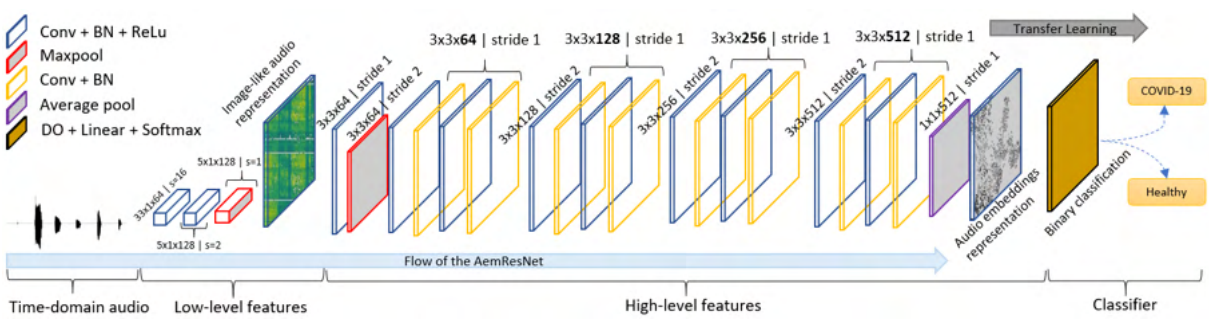

Fig. 1. AemResNet archictecture. The LLF and HLF blocks are pretrained with a large dataset to generate audio embeddings, and the classifier layer is trained by means of transfer learning.

correspondent to a spectral filterbank feature extraction. These signal handling values were chosen since we have observed heuristically that results in efficient development of CNNs for audio classification tasks. The sampling frequency of the audio signal is an important variable which can be set to a higher value that might result in a better audio quality, at the cost of increasing the model complexity (number of parameters and size). From this, we have found that 16 KHz is a good tradeoff between audio quality for classification and low complexity aimed to the purpose of deployment as mentioned before for e2e audio classification solutions.

The output of the LLF block creates an image-like tensor that is the direct input to the HLF block. The HLF block is built as a CNN architecture which is the most common approach for computational vision. For AemResNet, we set this HLF stage with a ResNet topology of 18 layers [13]. Details for this ResNet are also shown in Fig 1. The output of its last convolutional layer is average pooled to produce a vector of 512 audio embeddings that represent a condensed representation of the audio sample. This average pool layer brings flexibility when dealing with different lengths of audio inputs, while maintaining the same parameters of the architecture.

The last stage of the AemResNet acts as a classifier, which is the composition of a dropout layer (DO) to reduce overfitting and a fully connected layer with linear activation functions. At the last part of this block, a SoftMax layer is used at the output to present the normalized values based on the number of classes specified.

EXPERIMENTATION

AemResNet was pre-trained over a large set of audio data, this resulted in a pre-trained model that is later fine-tuned based on the audio classification task such as COVID-19 diagnosis based on respiratory sounds. All experimentation was executed using the Pytorch framework [14].

PRE-TRAINING STAGE

Both LLF and HLF stages are pre-trained using AudioSet, a large dataset of manually annotated audio events released by Google [15], containing 2.1 million samples equivalent to 5.8 thousand hours of recordings in which 527 different audio classes were labeled. Before using this embedding generator model for a specific classification application, the final classification block is removed, i.e. the fully connected layer, resulting in a 512-dimensional audio embeddings representation as the output. AemResNet used Audioset as pretraining as follows: the single channel raw audio is downsampled to 16 KHz, it is then standardized in amplitude by subtracting the mean and dividing it by the standard deviation of the signal. As well, data augmentation techniques such as random noise addition, random segment cropping of the audio sample, random gain variation and the widely used mixup data augmentation technique. During the training stage, a batch of audio clips were selected randomly into the form of mini batches to train the model. For validation, the complete standardized audio clips were used for inference.

Adam optimizer with a learning rate of $5x10^{-4}$ was used, with a weight decay of $1x10^{-8}$, and a mini batch size of 512 over 80 epochs. Cosine aligned learning rate schedule was used. This audio embedding was trained using the available unbalanced set and validated with the evaluation set for the 527 classes. This audio embedding generator model resulted in 11,744,143 number of trainable parameters, with a mean average precision (mAP) of 0.3690 over the AudioSet evaluation data, and it is the exact same one used in [12].

Task	Learning Rate	Learning Rate %	Dropout
Task 1	$1x10^{-3}$	80	0.2
Task 2	$1x10^{-3}$	60	0.2
Task 3	$1x10^{-6}$	90	0.9

Table I. Optimal hyper-parameters found for aemresnet per task.

END-TO-END CNN FOR COVID-19 DETECTION

The pre-trained audio embedding generator was used to train a COVID-19 classifier using the commonly adopted transfer learning technique [16]. For this purpose, the Cambridge Crowdsourced dataset described in [6] was used as the target application data. The University of Cambridge launched an application in Android and on a website [17] in which participants are asked to fill demographics general information and symptoms check. The dataset comprises 459 cough and breath samples from 378 users from Web and Android applications until May 2020. These data were annotated by experts and the audio samples were carefully checked to guarantee the quality of the data that contains only cough and breathing. As a preprocessing step, audio data was processed to be single channel with 16kHz sampling rate on a 16-bit resolution, and standardized in amplitude. Both web and Android app sources were used as samples for experimentation, and followed the authors proposal in [6] into three different experimental tasks:

- Task 1. Cough + breath sounds are used to classify COVID-19 vs healthy samples from 66 user (282 samples which represented 32% of the audio samples) and 220 users (596 samples representing 68% of the audio samples), respectively. Where COVID-19 samples included patients with and without cough or symptoms against healthy patients that have not reported symptoms as well as a clean medical history.
- Task 2. Cough sounds are used to classify COVID-19 vs healthy samples from 23 user (54 samples which represented 63% of the audio samples) and 29 users (32 samples representing 37% of the audio samples), respectively. Where COVID-19 samples included patients that reported cough as a symptom, and healthy patients that presented cough as well but have a clean medical history.
- Task 3. Breath sounds are used to classify healthy vs COVID-19 samples from 23 user (54 samples which represented 73% of the audio samples) and 18 users (20 samples representing 27% of the audio samples), respectively. Where COVID-19 samples included patients that reported cough as a symptom, and healthy patients that presented cough as well but have declared asthma in their medical history.

The training strategy followed for this application is similar to the pre-training of the audio embedding generator: Adam optimizer was used with a learning rate of $1x10^{-3}$ for Task 1 and Task 2, Task 3 used $1x10^{-6}$, weight decay of 1e-8, mini batch size of 32 over 400 epochs, cosine aligned learning rate schedule, and warm up of 20 epochs before mixup. It is important to notice that the Cambridge Crowdsourced dataset presents a highly imbalanced number of samples per condition, i.e. there is a significantly larger number of healthy breath and cough samples compared to the COVID-19 ones (approximately 73% against 27%, respectively.). Due to this issue, a focal loss approach was used in the loss function [18] for all of our experiments, resulting in a more efficient training process.

An exhaustive search was executed to find optimal learning rate and dropout values

		TASK 1			TASK 2			TASK 3		
Model	*Folds*	*Precision*	*Recall*	*F1*	*Precision*	*Recall*	*F1*	*Precision*	*Recall*	*F1*
AemResNet[a]	**5**	**0.6975**	**0.7235**	**0.7332**	**0.8697**	**0.8850**	**0.8773**	**0.9040**	**0.8300**	**0.8654**
SVM [16]	10	0.7200	0.6900	0.7047[b]	0.8000	0.7200	0.7579[b]	0.6900	0.6900	0.6900[b]
Ensemble [17]	3	-	0.7020	-	-	-	-	-	-	-
1D CNN [18]	-	-	-	0.9078	-	-	0.8926	-	-	0.8913
CI-ResNet[19]	10	-	0.7700	-	-	0.5350	-	-	0.7740	-

Table II. Experimental Validation results obtained as the average across 5 folds for AemResnet compared to other published Works. [a]Approach proposed in this work [b]F1-score computed with Equation (3).

hyperparameters in the classifier block; learning rates from $1x10^{-3}$ to $1x10^{-6}$ and drop out values from 0.1 to 0.9 where explored. Additionally, as observed in previous works [9], [12], we also explored the flexibility of the audio embeddings generator to dynamically adapt to the current target application by allowing to adjust its weights during training. Different learning rate values for the LLF and the HLF were utilized as a percentage of the fully connected layer learning rate; to fine-tune the model, this percentage was swept through different values from 10% to 100% in increments of 10%. The optimal values found for AemResNet across the three different tasks can be found in Table I.

Since there is no suggested official data split available for training/validation of the developed classification models, we randomly defined a set of 5 custom folds with a split 80% of the data for training, and 20% for validation (80/20 split). In all 5 folds, the proportion of available healthy and COVID-19 samples in maintained in both the training and validation split. The model obtained after training the healthy vs and COVID-19 classifier using the pre-trained audio embedding generator resulted in 11,473,282 which is 2.3% less parameters due to a smaller classifier block with only 2 outputs. Additionally, the number of multiply-accumulate operations (MACs) results in 1.84×10^{9}.

For a quantitative assessment of the performance of the proposed AemResNet model, Precision, Recall, and F1-score metrics were used for better understanding of our proposed implementation. These metrics are defined by:

$$\boldsymbol{Prec} = \frac{TP}{TP+FP} \ (1), \boldsymbol{Rec} = \frac{TP}{TP+FN} \ (2), \boldsymbol{F1} = 2 \times \frac{Prec \times Rec}{Prec + Rec} \ (3).$$

In Equations (1) and (2), the TP represents the true positives or the number of correctly classified breath and/or cough sounds into healthy or covid, FP represents an incorrect classification, and FN represents a miss classification. Finally, the computation of the F1-score comptured as in (3) to have a single metric that represents the performance of our model. The experimental results obtained based on the metrics defined above are presented in the following section.

RESULTS AND DISCUSSION

All experimental results obtained with our proposed AemResnet model implementation, using the custom 5-fold random 80/20 splits, are analyzed in this section. To efficiently increase the robustness in the detection of COVID-19 in respiratory sounds, we leveraged on the use of transfer learning for better performance. Table II presents the performance results of our approach averaged over the defined 5 folds, trained and validated for Task 1, Task 2, and Task 3; this table also shows how the performance obtained by the AemResNet compares to results reported in recent published works that benchmark over the same dataset [6]–

[8], [19]. Although these works present their results based on different metrics, we made an effort to consolidate and compare the performance of our approach as much as possible.

We computed the F1-Score from the SVM system in [6] based on the reported Precision and Recall and using Equation (3). From this, it can be observed that AemResNet presents a slightly better F1-Score of around 3.0% for Task 1, but this difference is more significant for Task 2 (almost 12.0%), and for Task 3 (> 17.0%). This suggest that AemResNet can generalize better for COVID-19 detection if only one type of respiratory sounds is considered, i.e., cough or breath sounds in separate models.

Looking at the Recall results, we can compare with the works presented in [6] and [8]. In this context, the Recall metric represents how accurate are the models at correctly classifying healthy and COVID-19 sounds. We found that AemResNet yields better positive classification accuracy in Task 2 (>16.5%) and Task 3 (>5.6%). However, this was not the case for Task 1, where AemResNet results in <4.6% Recall. Lastly, we compared our F1-score results to the ones reported in [19]various models of Artificial Intelligence (AI, where AemResNet felt short to the 1D CNN used in their work, particularly for Task 1 (~17.5%). A major difference here could be the use of efficient data augmentation procedures, which would suggest that handling of more data would be expected to be beneficial. We believe we could adopt this type of data augmentation to increase the robustness of our own e2e model and constitutes part of our ongoing research. Overall, the results obtained by AemResNet suggest that the use of the pre-trained deep audio embeddings applied to the task of COVID-19 detection is a robust, convenient, and competitive approach.

CONCLUSION

The experimental results presented in this work prove that AemResNet can be applied to classify breath and cough sounds into healthy or COVID-19 samples, with comparable results to the existing SOTA reported in the literature. The attractive characteristic of this e2e approach is that it avoids the need of additional pre-processing steps for feature extraction at the front-end, thus facilitating its portability into an inference engine. Through the use of pre-trained deep audio embeddings generator, a COVID-19 detection classifier model was build through transfer learning that achieved a F1-score of 0.7332 for cough and breath sounds combined, 0.8773 for cough sounds, and 0.8654 for breath sounds, over the 2020 Cambridge Crowdsourced dataset.

REFERENCES

1. "Coronavirus." https://www.who.int/westernpacific/health-topics/coronavirus (accessed Sep. 18, 2021).

2. A. Manzoor, Q. Pan, H. J. Khan, S. Siddeeq, H. M. A. Bhatti, and M. A. Wedagu, "Analysis and Detection of Lung Sounds Anomalies Based on NMA-RNN," in 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Dec. 2020, pp. 2498–2504. doi: 10.1109/BIBM49941.2020.9313197.

3. L. Pham, I. McLoughlin, H. Phan, M. Tran, T. Nguyen, and R. Palaniappan, "Robust Deep Learning Framework For Predicting Respiratory Anomalies and Diseases," in 2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC), Jul. 2020, pp. 164–167. doi: 10.1109/EMBC44109.2020.9175704.

4. L. D. Pham, H. Phan, R. Palaniappan, A. Mertins, and I. Mcloughlin, "CNN-MoE based framework for classification of respiratory anomalies and lung disease detection," IEEE J. Biomed. Health Inform., pp. 1–1, 2021, doi: 10.1109/JBHI.2021.3064237.

5. K. Minami, H. Lu, H. Kim, S. Mabu, Y. Hirano, and S. Kido, "Automatic Classification of Large-Scale Respiratory Sound Dataset Based on Convolutional Neural Network," in 2019 19th International Conference on Control, Automation and Systems (ICCAS), Oct. 2019, pp. 804–807. doi: 10.23919/ICCAS47443.2019.8971689.

6. C. Brown et al., "Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data," San Diego, p. 11.

7. H. Coppock, A. Gaskell, P. Tzirakis, A. Baird, L. Jones, and B. Schuller, "End-to-end convolutional neural network enables COVID-19 detection from breath and cough audio: a pilot study," BMJ Innov., vol. 7, no. 2, pp. 356–362, Apr. 2021, doi: 10.1136/bmjinnov-2021-000668.

8. M. A. Nessiem, M. M. Mohamed, H. Coppock, A. Gaskell, and B. W. Schuller, "Detecting COVID-19 from Breathing and Coughing Sounds using Deep Neural Networks," in 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS), Jun. 2021, pp. 183–188. doi: 10.1109/CBMS52027.2021.00069.

9. P. Lopez-Meyer, J. A. del Hoyo Ontiveros, H. Lu, and G. Stemmer, "Efficient End-to-End Audio Embeddings Generation for Audio Classification on Target Applications," in ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, Jun. 2021, pp. 601–605. doi: 10.1109/ICASSP39728.2021.9414229.

10. J. J. Huang and J. J. A. Leanos, "AclNet: efficient end-to-end audio classification CNN," ArXiv181106669 Cs Stat, Nov. 2018, Accessed: Aug. 17, 2021. [Online]. Available: http://arxiv.org/abs/1811.06669

11. J. Huang, H. Lu, P. Lopez Meyer, H. Cordourier, and J. Del Hoyo Ontiveros, "Acoustic Scene Classification Using Deep Learning-based Ensemble Averaging," in Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019), 2019, pp. 94–98. doi: 10.33682/8rd2-g787.

12. C. A. Galindo-Meza, P. Lopez-Meyer, and J. A. del Hoyo Ontiveros, "Classification of Respiration Sounds Using Deep Pre-trained Audio Embeddings," p. 4.

13. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," ArXiv151203385 Cs, Dec. 2015, Accessed: Apr. 18, 2021. [Online]. Available: http://arxiv.org/abs/1512.03385

14. A. Paszke, S. Gross, and F. Massa, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, pp. 8024--8035.

15. J. F. Gemmeke et al., "Audio Set: An ontology and human-labeled dataset for audio events," in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Mar. 2017, pp. 776–780. doi: 10.1109/ICASSP.2017.7952261.

16. J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," ArXiv14111792 Cs, Nov. 2014, Accessed: Sep. 18, 2021. [Online]. Available: http://arxiv.org/abs/1411.1792

17. "New app collects the sounds of COVID-19," University of Cambridge, Apr. 06, 2020. https://www.cam.ac.uk/research/news/new-app-collects-the-sounds-of-covid-19 (accessed Sep. 18, 2021).

18. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," ArXiv170802002 Cs, Feb. 2018, Accessed: Apr. 18, 2021. [Online]. Available: http://arxiv.org/abs/1708.02002

19. K. K. Lella, A. Pja, and Department of Computer Applications, NIT Tiruchirappalli, Tamil Nadu, India, "Automatic COVID-19 disease diagnosis using 1D convolutional neural network and augmentation with human respiratory sound based on parameters: cough, breath, and voice," AIMS Public Health, vol. 8, no. 2, pp. 240–264, 2021, doi: 10.3934/publichealth.2021019.