Acta Acustica 2022, **6**, 29

https://doi.org/10.1051/aacus/2022029

Available online at:
https://acta-acustica.edpsciences.org

SHORT COMMUNICATION OPEN ACCESS

Learning complementary representations via attention-based ensemble learning for cough-based COVID-19 recognition

Zhao Ren[1,2,a,*], Yi Chang[3,a], Wolfgang Nejdl[2], and Björn W. Schuller[1,3]

[1] Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, 86159 Augsburg, Germany
[2] L3S Research Center, Leibniz University Hannover, 30167 Hannover, Germany
[3] GLAM – Group on Language, Audio & Music, Imperial College London, SW7 2AZ London, UK

Received 6 August 2021, Accepted 5 July 2022

Abstract – Coughs sounds have shown promising as a potential marker for distinguishing COVID individuals from non-COVID ones. In this paper, we propose an attention-based ensemble learning approach to learn complementary representations from cough samples. Unlike most traditional schemes such as mere maxing or averaging, the proposed approach fairly considers the contribution of the representation generated by each single model. The attention mechanism is further investigated at the feature level and the decision level. Evaluated on the Track-1 test set of the DiCOVA challenge 2021, the experimental results demonstrate that the proposed feature-level attention-based ensemble learning achieves the best performance (AUC: 77.96%), resulting in an 8.05% improvement over the challenge baseline.

Keywords: COVID-19, Cough sound, Ensemble learning, Attention mechanism, Complementary representation

1 Introduction

Cough sounds from patients with different respiratory illnesses have been proven to have distinct latent features, which can be extracted and fed into machine learning models for recognition purpose [1]. In the COronaVIrus Disease 2019 (COVID-19) pandemic, coughing is one of the main modes of COVID-19 dissemination [2]. It is worthwhile to investigate the feasibility of automatic cough-based COVID-19 recognition, as it is potentially cheaper and faster than the pre-existing diagnosis methods, e.g., polymerase chain reaction testing.

Deep learning tends to outperform traditional machine learning by learning highly non-linear transformations for disease detection [3]. However, most COVID-related cough sound databases [4, 5] are small-scale, making it challenging to train deep learning models. Transfer learning is promising to transfer the knowledge learnt from large-scale datasets to a new small dataset. Image-based models trained on ImageNet [6] were successfully employed for audio classification [3], and audio-based models trained on AudioSet [7] performed better than image-based models in [8]. In our study, both image-based and audio-based models, as well as feed-forward deep neural networks (DNNs), are applied and compared.

Ensemble learning takes full advantage of multiple models for better performance. For example, features extracted by histogram-oriented gradient and a Convolutional Neural Network (CNN) model were simply concatenated for X-ray-based COVID-19 recognition [9]. Either majority voting or calculating the max/average score of the predicted probabilities was employed to fuse the predictions at the decision level [10]. However, for the conventional fusion methods, there is a difficulty in integrating each model's contribution into the final results when fusing multiple models. In this regard, weighted fusion was used to sum up all features or predictions with weight values [11]. A neural network was trained to fuse multiple representations [12], and an attention-based fusion was attempted to learn the weights of each model [13]. Nevertheless, few studies have investigated and compared feature-level and decision-level attention-based fusions.

To this end, for the first time we propose assembling multiple cough-based COVID-19 recognition models (i.e., a feed-forward DNN model, an image-based model, and an audio-based model) with feature-/decision-level attention, by assuming attention can estimate each feature item's or each prediction's contribution to complementing multiple representations. Key results demonstrate the attention-based ensemble learning effectively outperforms single models and other conventional fusion methods.

*Corresponding author: zren@l3s.de
[a] Equal contribution.

2 Methodology

In our study, three single-model representations are learnt from hand-crafted features, colourful log Mel spectrogram images, and original log Mel spectrograms, respectively. The attention-based ensemble learning is further proposed to extract helpful information from each representation.

2.1 Single-model representations

From the perspective of the input data format of neural networks, three single-model representations are extracted.

Hand-crafted-feature-based representations. To explore the performance of hand-crafted features, three feature sets are extracted, including a log Mel feature set, a Mel Frequency Cepstral Coefficients (MFCC) feature set, and a Computational Paralinguistics ChallengE (ComParE) feature set [14]. The log Mel and MFCC feature sets respectively calculate 26 Mel bins and 14 MFCCs for Low-Level Descriptors (LLDs). For those LLDs, we apply 100 functionals from the ComParE feature set, resulting in 2600 log Mel features and 1400 features for each audio wave. The ComParE feature set generates 6373 features for each audio signal. A feed-forward DNN model is used to process the hand-crafted features. The activations from the intermediate Fully Connected (FC) layers are further extracted as the hand-crafted-feature-based representations for the later ensemble learning.

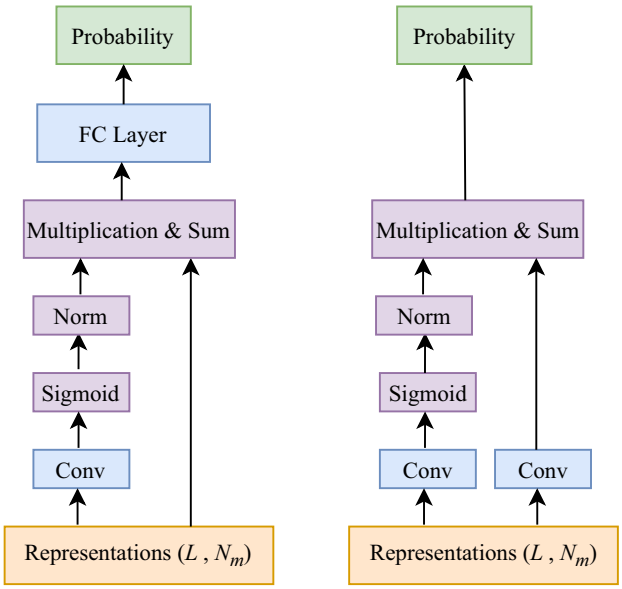
Deep Image-from-audio-based representations. Two typical CNN models pre-trained on ImageNet [6], VGG11 [15] and ResNet34 [16], are used to extract features from colourful log Mel spectrogram images [17] which have three channels as those of the original image inputs of the pre-trained models. Both models are fine-tuned with added FC layers for the final results. Similar to the hand-crafted-feature-based representations, the deep image-from-audio-based representations are extracted from the added FC layers.

Deep audio-based representations. The audio-based features herein are extracted by pre-trained models learnt from AudioSet [7]. Both CNN14_16k and ResNet38 models [18] are fine-tuned with added FC layers on the log Mel spectrograms. Similarly, the deep audio-based representations are extracted from these extra FC layers.

Although deep audio-based representations outperformed deep image-from-audio-based representations by mitigating the gap between natural images and time-frequence representations of audio waves in [8], we assume there is hidden difference between the two representations. Therefore, both of them are used in this study.

2.2 Attention-based ensemble learning

The representations to be fused are defined as a tensor R with a shape of (L, N_m), where L is each representation's length, and N_m denotes the number of the representations.



(a) Feature-level Attention (b) Decision-level Attention

Figure 1. The attention-based ensemble learning.

2.2.1 Feature-level attention

From the N_m representations, attention-based feature-level fusion intends to learn a new vector based on the contribution of each feature item (Fig. 1a). A one-Dimensional (1D) convolutional layer with a kernel size of 1 and an output channel number of L, followed by a sigmoid function, is applied to the tensor R, outputting a new tensor A^f with the same shape of R. Afterwards, A^f is normalised and multiplied with R using

$$M^f = \frac{A^f}{\sum_{j=1}^{N_m} A_j^f} \odot R, \tag{1}$$

where M^f is the element-wise multiplication result. The normalised A^f is considered as the contribution (i.e., weight) for each feature item $R_{i,j}$, $i \in [1; L]$, $j \in [1; N_m]$. M^f is then summed up along the axis across multiple representations to generate a vector, which is fed into an FC layer to compute the final probability. To enhance the flexibility of the feature-level attention, the output channel of the convolutional layer could be a different number L'. In this way, an additional convolutional layer with an output channel number L' should be added to process R so that the outputs from the two convolutional layers have the same size (L', N_m).

2.2.2 Decision-level Attention

The decision-level attention processes R with two 1D convolutional layers with a kernel size of 1 and an output channel number of 1, generating two new vectors (Fig. 1b). One of the convolutional layers, followed by a

sigmoid function, takes the original tensor R and outputs a vector A^{d1} with a length of N_m. Next, A^{d1} is normalised and multiplied with the other newly generated vector A^{d2} by

$$M^d = \frac{A^{d1}}{\sum_{j=1}^{N_m} A_j^{d1}} \odot A^{d2}, \quad (2)$$

where M^d is the element-wise multiplication result. Afterwards, M^d is summed up along the axis across multiple representations for the ultimate probability value. Compared to the feature-level attention, the decision-level attention learns the weight values for each representation rather than each feature item, as the output channel number of the convolutional layers is related to the class number. With fewer weight values to learn, the decision-level attention is coarser-grained than the feature-level one.

3 Experimental results

3.1 Database

The Track-1 dataset of the DiCOVA challenge 2021 [5] consists of 1040 cough sounds recorded from 1040 subjects (non-COVID: 965, COVID: 75). Five train-validation folds are split from the dataset, leading to 772 non-COVID and 50 COVID samples in each training set, and 193/25 non-COVID/COVID samples in each validation set. Moreover, an additional blind test set consists of 233 audio samples. All cough sound recordings are sampled into 44.1 kHz and stored in .FLAC format. As for performance evaluation, the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) [5] are assessed. Additionally, the specificity is calculated at 80% sensitivity.

3.2 Experimental setup

We re-sample all audio recordings into 16 kHz. To generate 224-frame log Mel spectrograms as required by the image-based models, the audio samples are segmented by a non-overlapping sliding window with a time length of 57, 600 frames. The window length and the overlap during short-term Fourier transformation are set as 512 and 256. We adjust the number of Mel bins as 64 and 128 to align with the input dimension requirements of the pre-trained audio-based and the image-based models. The log Mel spectrograms are converted into images with the "jet" colourmap for the image-based models. The mixup approach [18] is utilised to augment the extracted hand-crafted features and (colourful) log Mel spectrograms 1.5 times, respectively.

For the hand-crafted-feature-based representations, the feed-forward DNN models consist of three FC layers with neurons' number 1024, 256, and 1. For the other two types of representations, three trainable FC layers with the same number of neurons are added after the pre-trained single models. All representations are extracted from the respective second FC layer. The layers before the inherent FC layers of the pre-trained models (the second FC layer of VGG11, the first FC layers of ResNet34, ResNet38, and CNN14_16k) are frozen, and the others are removed. Specifically, the final convolutional block of ResNet34 is set to be trainable, since the transferability of its unique FC layer is limited. During training, the model parameters are updated within 30 epochs and a batch size of 16. The "Adam" optimiser with an initial learning rate of 0.001 is experientially chosen, and the learning rate decays by 0.1 after every 10 epochs for a stable training process. We apply the binary cross-entropy with logits loss as the loss function.

For experimental comparison, we fuse the representations by max and average fusions at the feature/decision level. The feature-level max and average fusions respectively calculates the maximum and average values across multiple 256-dimensional representations from the second last FC layers, while the decision-level ones computes the maximum and average predicted probabilities. The inputs of both feature-level and decision-level attentions are the outputs from the second last FC layers.

3.3 Results and discussions

In Table 1, the reason for the better results on the test set than those on the validation sets in most cases is probably all 1040 cough sounds are used for training a model which is verified on the test set instead of smaller training sets during cross validation. Both log Mel and ComParE features perform well on the test set, indicating it is promising to extract log Mel spectrograms as the CNNs' inputs. When comparing the performance of log Mel features and MFCC features, log Mel features perform better than MFCC features on the test set, perhaps because MFCC features are highly compressible and 14 coefficients are not enough to represent the characteristics of COVID cough sounds. The deep audio-based representations mostly outperform the deep image-from-audio-based representations, perhaps because audio-based models are more suitable for audio-related tasks than image-based models. The best three-type representations are from the feed-forward DNN model on the ComParE features, VGG11 on the colourful log Mel spectrogram images, and ResNet38 on the log Mel spectrograms, respectively.

The above representations are further processed by ensemble learning. Compared to the official baseline (i.e., average validation AUC: 68.81%, test AUC: 69.91%), the results of the three single-model representations are similar or slightly inferior on the validation and test sets. The ensemble learning approaches with complementary single-model representations ameliorate most single models and mostly outperform the baseline system. At both the feature and decision levels, the average fusion performs better than the max fusion, perhaps because max fusion neglects beneficial representations when selecting the maximum values only. Remarkably, the feature-/decision-level attention-based fusion outperforms both max and average fusions, indicating attention-based fusion can better complement multiple representations. Notably, the feature-level attention obtains the best AUC of 77.96% on the test set, which

Table 1. Performances of the proposed approaches on the DiCOVA Track-1 database. The performances on the validation sets are averaged over the five folds. SD = standard deviation.

	Validation		Test	
(%)	Sp (SD)	AUC (SD)	Sp	AUC
Baseline [5]	–	68.81	–	69.91
Hand-crafted-feature-based representations				
Log Mel	33.16 (8.49)	60.67 (6.15)	27.60	67.41
MFCC	30.88 (7.63)	61.16 (3.65)	23.44	55.63
ComParE	33.58 (8.08)	65.50 (1.96)	40.10	66.90
Deep image-from-audio-based representations				
VGG11	32.23 (2.73)	63.09 (2.55)	42.71	65.69
ResNet34	20.41 (3.83)	51.69 (3.80)	38.02	58.75
Deep audio-based representations				
CNN14_16k	42.28 (5.49)	67.82 (3.29)	44.27	67.77
ResNet38	39.27 (14.81)	68.36 (4.54)	45.31	65.66
Ensemble learning				
Feature-max	00.00 (0.00)	63.73 (3.94)	28.12	65.56
Feature-avg	35.54 (8.59)	68.51 (2.09)	61.46	73.72
Feature-attention	44.56 (6.29)	70.56 (3.01)	**59.38**	**77.96**
Decision-max	31.40 (8.20)	66.66 (2.25)	38.02	66.44
Decision-avg	35.03 (9.67)	68.57 (2.98)	55.73	73.25
Decision-attention	39.17 (9.41)	68.21 (3.92)	**59.38**	**77.36**

Bold values: The feature-level attention (specificity: 59.38%, AUC: 77.96%) outperforms the other two feature-level fusions (i.e. feature-max and feature-average (avg)), and the decision-level attention (specificity: 59.38%, AUC: 77.36%) performs the best among the three decision-level fusions.

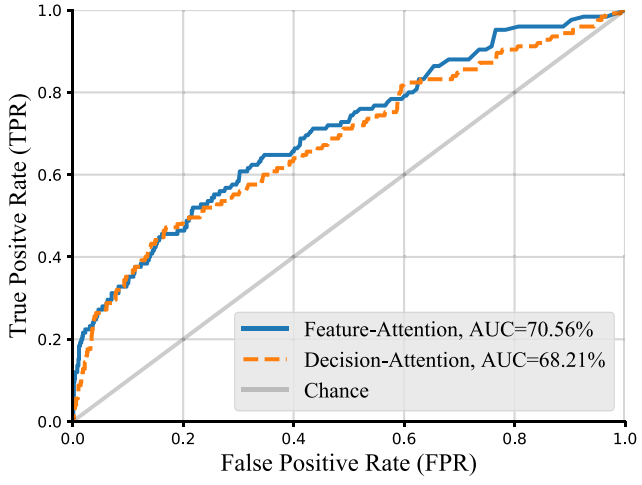


Figure 2. The average ROC curves of the attention-based ensemble learning on the validation sets.

is a significant amelioration of the baseline system's performance ($p < 0.05$ by a one-tailed z-test).

Apart from AUC values, the specificity results are also compare in Table 1. Higher specificity is corresponding to lower false positive rate. Both attention-based fusion models perform with a high specificity of 59.38% on the test set. The feature-level average fusion has the highest specificity (61.46%) while it perform slightly worse than the attention-based fusions on AUC (73.72%), perhaps because its ROC curve is not stable on all thresholds. To analyse the two attention-based fusion methods, the average Receiver Operating Characteristic (ROC) curves over the five-fold validation sets are depicted in Figure 2. We can see that both attention-based fusion methods yield finer True Positive Rates (TPRs) at given False Positive Rates (FPRs) than the chance. The ROC curves illustrate that the attention-based fusion approaches are adequate to recognise COVID-19 from cough sounds.

4 Conclusions and future work

Three single-model representations were extracted in this work, i.e., hand-crafted-feature-based representations, deep image-from-audio-based representations, and deep audio-based representations. The proposed attention-based ensemble learning was further applied to learn these complementary representations. On the DiCOVA challenge 2021 Track-1 database, both feature- and decision-level attention-based fusions outperformed the single-model classifiers and the max/average fusion for COVID-19 recognition. In future efforts, we will augment the training data by using more COVID-19-related databases and more data augmentation methods, e.g., SpecAugment [18]. The two attention-based fusion mechanisms will be further compared and analysed on other acoustic tasks, such as speech emotion recognition [19].

Conflict of interest

The authors declare no conflict of interest.

Acknowledgments

This work was partially supported by the BMBF project LeibnizKILabor with grant No. 01DD20003, and the Horizon H2020 Marie Skõõlodowska-Curie Actions Initial Training Network European Training Network (MSCA-ITN-ETN) project under grant agreement No. 766287 (TAPAS). The authors would thank the friendly discussions from their colleagues Lukas Stappen and Vincent Karas.

References

1. C. Infante, D. Chamberlain, R. Fletcher, Y. Thorat, R. Kodgule: Use of cough sounds for diagnosis and screening of pulmonary disease, in Proc. GHTC, San Jose, CA, 2017, 1–10.
2. A. Amit, R. Bhardwaj: Reducing chances of COVID-19 infection by a cough cloud in a closed space. Physics of Fluids 32 (2020) 101704.
3. Z. Ren, N. Cummins, V. Pandit, J. Han, K. Qian, B. Schuller: Learning image-based representations for heart sound classification, in Proc. DH, Lyon, France, 2018, 143–147.

4. C. Brown, J. Chauhan, A. Grammenos, J. Han, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, C. Mascolo: Exploring automatic diagnosis of COVID-19 from crowdsourced respiratory sound data, in Proc. ACM SIGKDD, virtual event, 2020, 3474–3484.
5. A. Muguli, L. Pinto, N. Sharma, P. Krishnan, P.K. Ghosh, R. Kumar, S. Bhat, S.R. Chetupalli, S. Ganapathy, S. Ramoji, V. Nanda: DiCOVA Challenge: Dataset, task, and baseline system for COVID-19 diagnosis using acoustics, in Proc. INTERSPEECH, Brno, Czech Republic, 2021, 901–905.
6. J. Deng, W. Dong, R. Socher, L. Li, K. Li, L. Fei-Fei: ImageNet: A large-scale hierarchical image database, in Proc. CVPR, Miami, FL, 2009, 248–255.
7. J.F. Gemmeke, D.P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R.C. Moore, M. Plakal, M. Ritter: Audio Set: An ontology and human-labeled dataset for audio events, in Proc. ICASSP, New Orleans, LA, 2017, 776–780.
8. T. Koike, K. Qian, Q. Kong, M.D. Plumbley, B. Schuller, Y. Yamamoto, Audio for audio is better? An investigation on transfer learning models for heart sound classification, in Proc. EMBC, Montreal, Canada, 2020, 74–77.
9. N. Alam, M. Ahsan, M.A. Based, J. Haider, M. Kowalski: COVID-19 detection from chest X-ray images using feature fusion and deep learning. Sensors 21 (2021) 1480.
10. S. Medjkoune, H. Mouchere, S. Petitrenaud, C. Viard-Gaudin: Handwritten and audio information fusion for mathematical symbol recognition, in Proc. ICDAR, Beijing, China, 2011, 379–383.
11. W. Wei, Q. Jia, Y. Feng, G. Chen: Emotion recognition based on weighted fusion strategy of multichannel physiological signals. Computational Intelligence and Neuroscience 2018 (2018) 5296523.
12. Y. Ma, Y. Hao, M. Chen, J. Chen, P. Lu, A. Kosir: Audio-visual emotion fusion (AVEF): A deep efficient weighted approach. Information Fusion 46 (2019) 184–192.
13. S. Chen, Q. Jin, Multi-modal conditional attention fusion for dimensional emotion prediction, in Proc. ACM Multimedia, Amsterdam, The Netherlands, 2016, 571–575.
14. F. Eyben, F. Weninger, F. Gross, B. Schuller: Recent developments in openSMILE, the Munich open-source multimedia feature extractor, in Proc. ACM Multimedia, Barcelona, Spain, 2013, 835–838.
15. K. Simonyan, A. Zisserman: Very deep convolutional networks for large-scale image recognition, in Proc. ICLR, San Diego, CA, 2015, 14.
16. K. He, X. Zhang, S. Ren, J. Sun: Deep residual learning for image recognition, in Proc. CVPR, Las Vegas, NV, 2016, 770–778.
17. S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird, B. Schuller: Snore sound classification using image-based deep spectrum features, in Proc. INTERSPEECH, Stockholm, Sweden, 2017, 3512–3516.
18. Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, M. Plumbley: PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020) 2880–2894.
19. J. Han, Z. Zhang, G. Keren, B. Schuller: Emotion recognition in speech with latent discriminative representations learning, Acta Acustica united with Acustica 104 (2018) 737–740.

Cite this article as: Ren Z. Chang Y. Nejdl W. & Schuller BW. 2022. Learning complementary representations via attention-based ensemble learning for cough-based COVID-19 recognition. Acta Acustica, **6**, 29.