A GLANCE-AND-GAZE NETWORK FOR RESPIRATORY SOUND CLASSIFICATION

Shuai Yu [1], *Yiwei Ding* [1], *Kun Qian* [2*], *Bin Hu* [2*], *Wei Li* [1,3], *and Björn W. Schuller*[4,5]

[1] School of Computer Science and Technology, Fudan University, China
[2] School of Medical Technology, Beijing Institute of Technology, China
[3] Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, China
[4]GLAM – Group on Language, Audio & Music, Imperial College London, UK
[5]Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
{qian, bh}@bit.edu.cn

ABSTRACT

A plethora of great successes has been achieved by the existing convolutional neural networks (CNN) for respiratory sound classification. Nevertheless, simultaneously capturing both the local and global features can never be an easy task due to the limitation of a CNN's structure. In this contribution, we propose a novel glance-and-gaze network to address the aforementioned issue. The glance block aims to learn global information, while the gaze block is responsible for learning local patterns and suppressing the noises that attenuates the final performance. In the proposed method, both the global and local information can be extracted. Moreover, the spectral and temporal representations can be learnt via a feature fusion module. Experimental results on the largest public respiratory sound database demonstrate that the proposed model outperforms the state-of-the-art methods.

Index Terms— Computer Audition, Digital Health, Respiratory Sound Classification, Glance-and-Gaze Network, Feature Fusion

1. INTRODUCTION

Respiratory sound classification plays an important role in the pulmonary pathology [1, 2, 3]. This task is for example assistive when examining cough sounds associated with COVID-19 [4, 5, 6]. Respiratory sounds can generally be classified as 'normal respiratory sounds' or 'abnormal respiratory sounds'. Normal respiratory sounds are when no pulmonary disease exists, while abnormal respiratory sounds are heard when a pulmonary disease is present. Deep Learning based respiratory sound classification methods have attracted great research attention in the field of respiratory sound classification [7, 8, 9, 10].

The first two authors contributed equally. This work was supported by the BIT Teli Young Fellow Program from the Beijing Institute of Technology and National Key R&D Program of China (2019YFC1711800), NSFC(62171138). *Corresponding authors*: Kun Qian and Bin Hu.

Despite the successes of prior works, existing CNN based methods cannot simultaneously leverage both global and local information due to structural limits [7, 11, 12]. A respiratory sound has wide frequency bands and long time spans. It is important to obtain both, global and local information. Although stacked CNNs can increase the receptive field, it needs too many layers to achieve that. And it also would be very difficult to train such a deep network for this task, considering the scarce of the training data. In addition, since respiratory sound is generated from the lungs, and the heart is very close to the lungs [13], heart sound is inevitably involved in the respiratory sound. Therefore, to suppress the interference of the noises is critical for respiratory sound classification.

To address the issues mentioned above, inspired by the effectiveness of CNN-based methods [7, 11, 14], we propose a novel glance-and-gaze network for respiratory sound classification. The glance module aims to obtain global information of the spectrogram. The gaze module is responsible for learning local patterns and suppressing the noises that interfere with the final prediction. In addition, we propose a feature fusion module to dynamically fuse the features that represent the relevant frequency bands and time steps. To be specific, for the glance module, we obtain the global information from the frequency and time axes. Based on extending the idea of a shrinkage network used in fault diagnosis [15, 16], we devise a gaze module to learn local patterns and suppress the noises with a soft-thresholding mechanism. Finally, we propose a feature fusion module to dynamically assign weights for the information learnt on the frequency and time axes.

Two technical contributions are made: i) we propose a novel glance-and-gaze network to capture both of global and local information in the spectrogram. To the best of our knowledge, there is no such work for respiratory sound classification in the literature; ii) we propose a feature fusion module to dynamically fuse the the information learnt on the frequency and time axes.

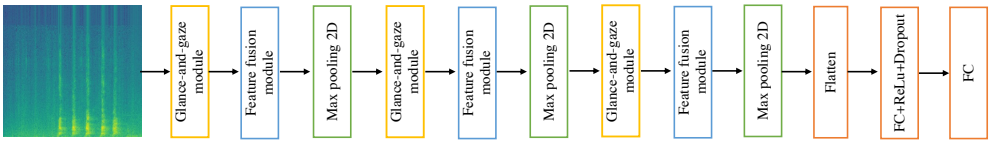


Fig. 1: The overall architecture of the proposed model. "FC" denotes a fully connected layer.

2. PROPOSED MODEL

The overall architecture of our proposed glance-and-gaze network is shown in Fig. 1. It has two modules: the glance-and-gaze module and the feature fusion module. In this network, we employ three glance-and-gaze modules, feature fusion modules, and max-pooling layers. The channel numbers are 32, 64, and 128 with forward propagation. The glance-and-gaze module and the feature fusion module are respectively addressed in this section.

2.1. Glance-and-gaze module

The detailed architecture of glance-and-gaze module is shown in Fig. 2. The left part of Fig. 2 is the glance part, the right part is the gaze part. The motivation of this module is to make the model learn both of the global and local information in the spectrogram.

2.1.1. Glance part

The glance part aims to learn global information from the frequency and time axis with fewer convolution layers. To achieve this, we first employ average pooling to compute the statistic feature descriptor of the spectrogram along the frequency or time axis. Then, we use a 1-D CNN as a weight operator to assign weights to the frequency bands/time step. The larger weights mean the frequency bands or the time step is important to the task.

Formally, given the input spectrogram or a feature map $\mathbf{S} \in \mathbb{R}^{F \times T}$, where F, and T denote the number of frequency bins and time steps, respectively. An average pooling is firstly applied to the spectrogram for calculating the distribution of magnitudes along the time axis. We use row average pooling to achieve this. The frequency descriptor $\mathbf{f} \in \mathbb{R}^{C \times F}$ can be calculated as:

$$\mathbf{f} = \frac{1}{T} \sum_{i<=T} s_{ij}, \tag{1}$$

where s_{ij} is the element in the i-th row and j-th column in $\mathbf{S}$. Similarly, we obtain the time descriptor $\mathbf{t} \in \mathbb{R}^{C \times T}$:

$$\mathbf{t} = \frac{1}{F} \sum_{j<=F} s_{ij}. \tag{2}$$

After we obtain the frequency and time descriptor, we use a 1-D CNN to assign weights to the frequency and time descriptor:

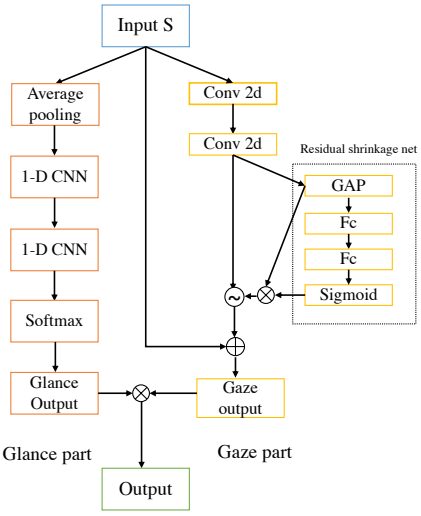


Fig. 2: The detailed architecture of the glance-and-gaze module.

$$\begin{aligned} \mathbf{w_f} &= Softmax(\mathbf{f} * \mathbf{v}), \\ \mathbf{w_t} &= Softmax(\mathbf{t} * \mathbf{v}), \end{aligned} \tag{3}$$

where $*$ denotes the convolution operation, and $\mathbf{v}$ is the kernel of the 1-D CNN. $\mathbf{w_f}$ and $\mathbf{w_t}$ are the weights of the frequency bands and time steps, respectively.

2.1.2. Gaze part

The gaze part aims to learn local patterns and suppress the noises in the spectrogram. To this end, we use two consecutive 2-D convolution layers with a kernel size of (3×3) and (5×5) to learn the local pattern. To suppress the noises in the spectrogram, we use a deep residual shrinkage network [15] to suppress the noises in the spectrogram. The detailed architecture is found in the right part of Fig. 2.

Formally, given the input feature map $\mathbf{S} \in \mathbb{R}^{C \times F \times T}$, where C denotes the number of channels. We apply two convolution layers above $\mathbf{S}$ to obtain a local feature map $\mathbf{S}'$:

$$\mathbf{S}' = Conv2d(Conv2d(\mathbf{S})). \tag{4}$$

Here, we omit the activation function and batch normalisation for simplification. After we obtain $\mathbf{S}'$, a residual shrinkage network [15] is used to suppress the noises. The rationale behind the network is to learn a soft threshold for denoising [16].

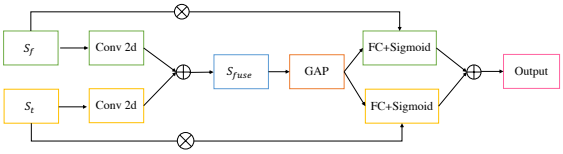


Fig. 3: The detailed architecture of the feature fusion module.

Specifically, we first employ global average pooling (GAP) to get the statistic value of each channel $\mathbf{g} \in \mathbb{R}^{C\times1\times1}$. Then, two fully connected (FC) layers and a *Sigmoid* layer are used to learn the importance of each channel $\mathbf{w}$:

$$\mathbf{w} = Sigmoid(FC(FC(\mathbf{g}))). \tag{5}$$

When we get the weight of each channel, we apply an element-wise multiplication between $\mathbf{w}$ and the feature map $\mathbf{S}'$ to learn the threshold for each channel $\mathbf{thre}$:

$$\mathbf{thre} = \mathbf{S}' \otimes \mathbf{w}. \tag{6}$$

After obtaining the $\mathbf{threshold}$, we apply a soft-thresholding operation to the feature map $\mathbf{S}'$:

$$\mathbf{S_{thre}} = \begin{cases} max(0, -S'_{cij} - thre_c) & S'_{cij} < 0 \\ max(0, \quad S'_{cij} - thre_c) & S'_{cij} >= 0, \end{cases} \tag{7}$$

where S'_{cij} is the element in the c-th channel, and $\mathbf{S_{thre}}$ is the resulting feature map.

Finally, we apply a residual operation to the feature map $\mathbf{S_{thre}}$ as the output of the gaze part:

$$\mathbf{S}'' = \mathbf{S} + \mathbf{S_{thre}}. \tag{8}$$

2.1.3. Glance-and-gaze output

Now, we have both of the global and local information given by Eq. 3 and Eq. 8; we combine them by a broadcast element-wise multiplication as the output of the glance-and-gaze module:

$$\begin{aligned} \mathbf{S_f} &= \mathbf{S}'' \otimes \mathbf{w_f}, \\ \mathbf{S_t} &= \mathbf{S}'' \otimes \mathbf{w_t}, \end{aligned} \tag{9}$$

where $\mathbf{S_f}$ and $\mathbf{S_t}$ represent the relevant frequency bands and time steps for respiratory sound classification.

2.2. Feature fusion module

To fuse the features that represent the relevant frequency bands and time steps, we propose a feature fusion module (FFM) to fuse them. The detailed architecture is shown in Fig. 3. First, we employ two parallel 2-D CNN layers with kernels (1×1) to equalise the number of channels of $\mathbf{S_f}$ and $\mathbf{S_t}$. Then, we fuse the two feature maps with element-wise addition. Since the fused feature map contains two types of information (i. e., frequency and time information), we need to re-estimate the importance of them. To this end, we use global average pooling and two FC layers to learn the weights of each channel. Finally, multiplications are performed between inputs and learnt weights to obtain re-weighted feature maps.

Method	Acc.	Sen.	Spe.	Pre.	Rec.	F1
w/o glance	81.4	81.4	81.3	80.1	86.5	83.2
w/o gaze	82.0	**87.8**	76.1	79.7	88.4	83.8
w/o FFM	82.1	83.9	80.3	83.0	88.0	85.4
Proposed	**84.7**	84.5	**84.9**	**84.4**	**89.0**	**86.6**

Table 1: Ablation study results of the proposed model on the ICBHI dataset.

3. EXPERIMENTS

3.1. Experiment setup

The International Conference in Biomedical and Health Informatics (ICBHI) 2017 database [17] used in this paper is the largest publicly available respiratory sound dataset. It consists of a total of 5.5 hours of recordings containing 6 898 respiratory cycles, of which 1 864 contain crackles, 886 contain wheezes, 506 contain both crackles and wheezes and the rest are normal, in 920 annotated audio samples from 126 subjects [17]. To evaluate the effectiveness of our proposed method, following [9, 18], we split the dataset into a training (70 %), and a testing (30 %) set. Following the convention [19], we use accuracy, sensitivity, specificity, precision, recall, and F1 as evaluation metrics. For pre-processing, we first resample all of the recordings to 22050 Hz. To analyse the recordings in the spectral domain, we compute the Mel-spectogram using a 2048-sample window size and a 512-sample hop size for all of the recordings in the dataset. Due to the imbalanced distribution of the training data, we perform simple augmentation on the training set. For the wheezes and wheezes & crackles recordings, we augment the recordings by changing the speed of the recordings as in [20].

3.2. Ablation study

To investigate how much of the proposed glance-and-gaze module contributes to this task, we first remove the glance part, then, the model has a gaze part and a feature fusion module. Because the feature fusion module is used to fuse features that represent relevant frequency bands and time steps, there is no need to fuse these features. As Table 1 shows, when focusing on accuracy, the performance is decreased by 3.90 %.

We then remove the gaze part: the model now has a glance part and a feature fusion module. Since the output of the glance part is the weight vectors of the frequency bands and time steps, we perform an element-wise multiplication between the input S and glance part to generate a feature map

Method	Acc.	Sen.	Spe.	Pre.	Rec.	F1
VGG16	65.3	54.6	76.0	63.1	63.2	63.1
CRNN	81.6	**85.5**	77.7	80.0	83.5	81.7
BiGRU+ATT	76.6	77.3	76.0	75.9	83.3	79.4
Proposed	**84.7**	84.5	**84.9**	**84.4**	**89.0**	**86.6**

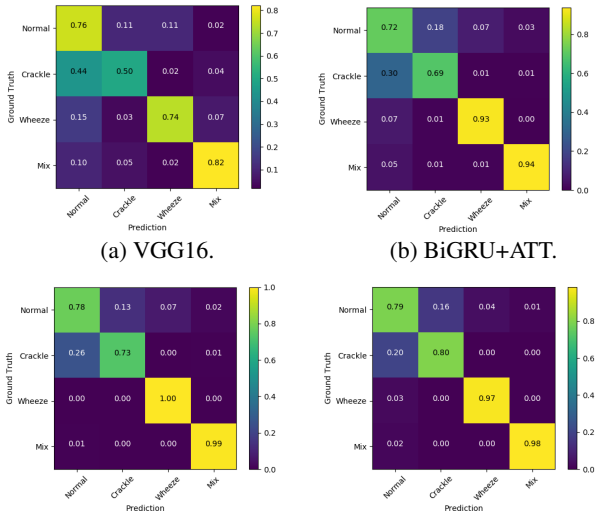
Table 2: Results of the proposed and baseline methods on the ICBHI dataset. The values in the table are percentiles.

for prediction. When focusing on accuracy, the performance is decreased by 3.19 %. However, when focusing on sensitivity, the ablated model achieves a better result than the proposed model. Analysing the result, we find that without the gaze part, the models are prone to predict a recording as an abnormal one as the local pattern is not properly studied. To investigate the effectiveness of the feature fusion module (FFM), we also remove this module. We replace it with a simple elment-wise addition. When focusing on accuracy, the performance is decreased by 3.07 %. The results show that the proposed feature fusion module improves the performance of this task.

3.3. Comparision with the state-of-the-art

The performance of the proposed model and the baseline methods are listed in the Table 2. Three baseline methods are compared in Table 2, including a pretrained VGG16, a CRNN [20] and a BiGRU+ATT [9]. We carefully tune the hyper-parameters of the three baseline methods to ensure that they reach their peak performances on our training dataset. The proposed model and the three baseline methods are trained on the same dataset. Compared with the baseline methods, the proposed method achieves the highest score in general. The results clearly confirm the effectiveness and robustness of our proposed model.

For comparison with other baselines, when focusing on accuracy, the proposed method outperforms the VGG16 by 29.7 %, the CRNN by 3.66 %, and the BiGRU+ATT by 10.57 %. However, when focusing on sensitivity, the proposed method achieves a comparable result as compared to the CRNN. We guess that this will be improved by a hierarchical classification model which first predicts whether the recording is normal or abnormal, then classifies whether the recording is a crackle or wheeze, or both. We leave this as a future research topic. We also use STFT features as input to the proposed model; the performance is decreased by $\sim$10 %. To further investigate what types of errors are fixed by the proposed model, we visualise the confusion matrix of the proposed model and baseline methods in Fig. 4. We can observe that our model achieves the highest accuracy in the prediction of the normal and crackle cases. When predicting wheeze and mix (i.e., crackle & wheeze) cases, our model achieves the second highest accuracy. The performance gains can be attributed to solving the normal and crackle errors.



(a) VGG16. (b) BiGRU+ATT.

(c) CRNN model. (d) Proposed model.

Fig. 4: Confusion matrix of the proposed model and baseline methods.

4. RELATED WORKS

In this section, we introduce the research related to the shown respiratory sound classification. Deep learning based methods have attracted great attention on this task. Ma et al. [7] proposed a CNN based bi-ResNet with short-time Fourier transformation (STFT) and wavelet features for respiratory sound classification. Demir et al. [11] employed a CNN model with parallel pooling structure to learn deep representations for LDA classifiers. RNN based models are also exploited: Kochetov et al. [21] proposed a gated recurrent unit (GRU) based model to mask the noises in the respiratory sounds. A long-short term memory (LSTM) model was also proposed to learn temporal relationship for this task [10]. Researchers also explored combinations of CNN and RNN. For example, Acharya et al. [20] proposed a CRNN model using Mel-spectrograms and achieved promising results.

5. CONCLUSION

In this paper, we proposed a glance-and-gaze network to leverage both local and global information for respiratory sound classification, which contains glance-and-gaze modules and a feature fusion module. Experimental results show that the proposed method outperforms three baseline methods on the ICBHI dataset. Designing a hierarchical classification model for respiratory sound classification will be a next step.

6. REFERENCES

[1] Noam Gavriely, Yoram Palti, Gideon Alroy, and James B Grotberg, “Measurement and theory of wheez-

ing breath sounds," *Journal of Applied Physiology*, vol. 57, no. 2, pp. 481–492, 1984.

[2] P Piirila and AR Sovijarvi, "Crackles: recording, analysis and clinical significance," *European Respiratory Journal*, vol. 8, no. 12, pp. 2139–2148, 1995.

[3] Kun Qian, Christoph Janott, Vedhas Pandit, Zixing Zhang, Clemens Heiser, Winfried Hohenhorst, Michael Herzog, Werner Hemmert, and Björn W Schuller, "Classification of the excitation location of snore sounds in the upper airway by acoustic multi-feature analysis," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 8, pp. 1731–1741, 2017.

[4] Jordi Laguarta, Ferran Hueto, and Brian Subirana, "COVID-19 artificial intelligence diagnosis using only cough recordings," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 1, pp. 275–281, 2020.

[5] Kun Qian, Maximilian Schmitt, Huaiyuan Zheng, Tomoya Koike, Jing Han, Juan Liu, Wei Ji, Junjun Duan, Meishu Song, Zijiang Yang, et al., "Computer audition for fighting the SARS-CoV-2 Corona crisis – Introducing the multi-task speech corpus for COVID-19," *IEEE Internet of Things Journal*, pp. 1–12, 2021.

[6] Kranthi Kumar Lella and PJA Alphonse, "A literature review on COVID-19 disease diagnosis from respiratory sound data," *AIMS Bioengineering*, vol. 8, no. 2, pp. 140–153, 2021.

[7] Yi Ma, Xinzi Xu, Qing Yu, Yuhang Zhang, Yongfu Li, Jian Zhao, and Guoxing Wang, "Lungbrn: A smart digital stethoscope for detecting respiratory disease using bi-resnet deep learning algorithm," in *Proc. BioCAS*, 2019, pp. 1–4.

[8] Wenjie Song, Jiqing Han, and Hongwei Song, "Contrastive embeddind learning method for respiratory sound classification," in *Proc. ICASSP*, 2021, pp. 1275–1279.

[9] Xuesong Zhao, Yanbo Shao, Juanyun Mai, Airu Yin, and Sihan Xu, "Respiratory sound classification based on BiGRU-Attention network with XGBoost," in *Proc. BIBM*, 2020, pp. 915–920.

[10] Diego Perna and Andrea Tagarelli, "Deep auscultation: Predicting respiratory anomalies and diseases via recurrent neural networks," in *Proc. CBMS*, 2019, pp. 50–55.

[11] Fatih Demir, Aras Masood Ismael, and Abdulkadir Sengur, "Classification of lung sounds with cnn model using parallel pooling structure," *IEEE Access*, vol. 8, pp. 105376–105383, 2020.

[12] Samiul Based Shuvo, Shams Nafisa Ali, Soham Irtiza Swapnil, Taufiq Hasan, and Mohammed Imamul Hassan Bhuiyan, "A lightweight CNN model for detecting respiratory diseases from lung auscultation sounds using emd-cwt-based hybrid scalogram," *IEEE J. Biomed. Health Informatics*, vol. 25, no. 7, pp. 2595–2603, 2021.

[13] E Saatci and A Akan, "Heart sound reduction in lung sounds by spectrogram," in *Proc. IFMBE*, Singapore, 2005, pp. 1727–1983.

[14] Shuai Yu, Xiaoheng Sun, Yi Yu, and Wei Li, "Frequency-temporal attention network for singing melody extraction," in *Proc. ICASSP*. 2021, pp. 251–255, IEEE.

[15] Minghang Zhao, Shisheng Zhong, Xuyun Fu, Baoping Tang, and Michael G. Pecht, "Deep residual shrinkage networks for fault diagnosis," *IEEE Trans. Ind. Informatics*, vol. 16, no. 7, pp. 4681–4690, 2020.

[16] David L Donoho, "Denoising by soft-thresholding," *IEEE transactions on information theory*, vol. 41, no. 3, pp. 613–627, 1995.

[17] BM Rocha, Dimitris Filos, L Mendes, I Vogiatzis, E Perantoni, E Kaimakamis, P Natsiavas, Ana Oliveira, C Jácome, A Marques, et al., "A respiratory sound database for the development of automated classification," in *Proc. ICBHI*, 2017, pp. 33–37.

[18] Hai Chen, Xiaochen Yuan, Zhiyuan Pei, Mianjie Li, and Jianqing Li, "Triple-classification of respiratory sounds using optimized s-transform and deep residual networks," *IEEE Access*, vol. 7, pp. 32845–32852, 2019.

[19] Bruno M Rocha, Dimitris Filos, Luís Mendes, Gorkem Serbes, Sezer Ulukaya, Yasemin P Kahya, Nikša Jakovljevic, Tatjana L Turukalo, Ioannis M Vogiatzis, Eleni Perantoni, et al., "An open access database for the evaluation of respiratory sound classification algorithms," *Physiological measurement*, vol. 40, no. 3, pp. 035001, 2019.

[20] Jyotibdha Acharya and Arindam Basu, "Deep neural network for respiratory sound classification in wearable devices enabled by patient specific model tuning," *IEEE transactions on biomedical circuits and systems*, vol. 14, no. 3, pp. 535–544, 2020.

[21] Kirill Kochetov, Evgeny Putin, Maksim Balashov, Andrey Filchenkov, and Anatoly Shalyto, "Noise masking recurrent neural network for respiratory sound classification," in *Proc. ICANN*, 2018, pp. 208–217.