Discovery RNNs, explainable RNN saliency visualization, and its application to unsupervised segmentation of COVID-19 forced coughs

Ferran Hueto[1,2]*, Prithvi Rajasekaran*[1]*, Jordi Laguarta*[1]*, Sanjay Sarma*[1]*, Brian Subirana*[1,2]

[1]MIT AutoID Lab, 77 Massachusetts Avenue, Cambridge MA 02139. USA
[2]Harvard University, Massachusetts Hall, Cambridge, MA 02138. USA

subirana@mit.edu

Abstract

Deep learning and Recurrent Neural Networks (RNNs) have revolutionized the field of audio analysis, providing flexible and robust recognition and categorization. Their wide adoption in hospitals and medical institutions is nonetheless still limited, in part because of their lack of explainability and possibility for expert cognitive knowledge. In this paper, we introduce Discovery RNNs, a neural network based architecture that learns to generate saliency visualizations by segmenting areas of interest in audio samples, allowing to better understand how predictions are made by target models and visualize unseen samples to assess the risk of errors. We demonstrate our approach on a forced cough sample dataset for COVID-19 screening, with two targeted models for the binary problems of gender and spoken language. Target model accuracies, 67.6% and 78.6% respectively, demonstrated for the first time that forced coughs can be used to extract cultural and biological information. Our results show that our visualization approach is effective in identifying areas of interest, indirectly segmenting cough events within the samples, while revealing a bias towards padded samples in the dataset. More work in this direction could enable the transfer of algorithmic knowledge to physicians by harnessing model input visualizations for educational purposes.

Index Terms: recurrent neural networks, autoencoders, cough analysis, deep learning explainability, audio segmentation.

1. Introduction

Deep learning algorithms generate predictions on a problem by adjusting a set of weights and biases to approximate a continuous function based on a set of training examples - when these examples are balanced, the algorithm is usually able to generalize to other similar examples [1], but it is often hard to explain when and why. In the field of audio processing, Recurrent Neural Networks (RNNs) have become the new golden standard, boasting state-of-the-art results in multiple tasks such as speech recognition [2] or audio encoding [3]. Although some work has been done on improving the explainability of deep learning models through visualizations [4, 5], this has seldomly been applied to RNNs directly. This is especially relevant for high consequence applications where quantifying the risk and attributing responsibility from a missed prediction prevent the wide implementation of deep learning algorithms - as is the case in medicine [6] despite other super-human examples of deep learning in dermatology [7], [8], psychiatry [9], gynecology[8], cognitive impairment [10, 11] and ophthalmology[12]. Gaining a better understanding of the inner workings of these algorithms could prove instrumental in their deployment, specifically in diagnosis and screening applications.

In this paper, we focus on the task of cough analysis. Medical doctors use different techniques of lung and trachea auscultation, one of which being cough analysis, for assessment and diagnosis of patients [13, 14]. Although effective, these techniques do not go beyond a rather superficial analysis, classifying spontaneous coughs as wet or dry, and finally formulating a diagnosis when contrasted with other symptoms or patient information [15]. In a recent study, we suggested cough recordings could be used for diagnosis of COVID-19 [16]. COVID-19 is a virus which was classified by the WHO as a worldwide pandemic [17], and is characterized by a very high rate of infection, including between asymptomatic patients. Forced cough analysis, rather than spontaneous, could provide a non-invasive way to identify all patients, whether symptomatic or not. Nonetheless, when using a large scale raw forced cough dataset, we were unable to replicate our promising ResNet-based COVID-19 discrimination results of 80% classification accuracy [16]. Many potential reasons could underlay such change and the prospect of using RNNs to improve on our previous research led us to the development of a new architecture to discover what the actionable source of our model's lack of success was.

We introduce a novel architecture, Discovery RNNs, that generate visualizations of salient areas for Deep Convolutional Neural Networks (DCNNs) and Recurrent Neural Networks (RNNs) in a weakly-supervised manner. This architecture is composed of three components. The first is an RNN model that generates classification predictions on samples (see section 2.2). We then define in section 2.3 a Collaborative Attention Saliency model, inspired by adversarial training of GANs, which generates saliency masks by reducing the number of activated pixels in an input image while tracking the target RNN model's recognition rate . Finally, in section 2.4 we define a human feedback loop which enables expert cognitive bias to inform saliency visualizations. Our architecture does not only allow to generate salience maps of processed samples, but to make predictions on maps of unseen data. These generated images can serve as a visualization tool to further the explainability of the targeted models by identifying specific features important for the task at hand, or as an alternative to spot samples where the target model might be prone to make inaccurate predictions.

We evaluate our architecture on a forced coughs dataset, which was collected for the purpose of COVID-19 detection, in the tasks of gender and spoken language classification. We demonstrate for the first time that forced coughs can be used to extract cultural and biological information from a subject, and through visualizations, confirm the robustness and explainability of our algorithms.

There are two major goals when visualizing how deep neural networks make predictions. The first consists of identifying areas a model pays attention to in specific samples [18, 19], whereas the second goal is to find instead inputs that maximize specific classes (or outputs) to better understand what the algorithm is looking for [20, 21]. Our proposed architecture is

able to achieve both through hyperparameter optimization and expert cognitive bias. Approaches in the first domain include conditional feature sampling as portrayed in [4], or through class activation mapping in [5]. Bazzani et al. used a similar approach to ours, by selectively masking the input image of a model while tracking recognition rate with an agglomerative clustering algorithm [22], but lacked both an expert feedback loop, and the ability to generate saliency predictions on unseen samples. Approaches tackling the second goal include [23], which maximized activation of multifaceted neurons to visualize targeted features by their computer vision models. All of the reviewed approaches have a common focus on computer vision tasks, which use variations of CNNs, and none were implemented in the context of audio analysis or in RNNs.

Most approaches to model visualization for RNNs tackle comparative behavior between different memory gates (e.g. [24, 25]), visualizing how knowledge from previous timesteps is transferred to future states. Some weakly-supervised audio segmentation approaches are applicable to this topic as well. Keren et alt. used Siamese neural networks for audio event segmentation by computing cosine similarity of the sample's latent space with a reference event [26], and [27] performed CNN based event tagging from a related clip dataset. Although related, these failed to directly address the problem of RNN visualization and explainability, specifically to identify salient areas of input samples.

2. Methods

2.1. Cough Dataset

We collected variable length forced cough audio recordings (up to 15 seconds) through an internal open website tool (opensigma.mit.edu, approved by IRB 2004000133) for the purpose of COVID-19 detection [16], accompanied by a set of 10 multiple choice questions related to the diagnosis of the disease and general subject information (see figure 1 for a list of all data columns), with an estimated subject count of 35.000 and 100,000 samples. All samples were saved without compression in WAV format (16kbs bit-rate, single channel, opus codec). Suspicious data entries were identified through a 3 component Principal Component Analysis and discarded for manual analysis. For the purpose of this paper, we stripped all but 2 data categories from all samples: gender (male, female) and spoken language (English, Spanish). Figure 1b describes population distribution for these labels. Individual cough events for a small subset of the test dataset were manually identified for the purpose of visualization in figure 4.

Although Mel Frequency Cepstral Coefficients (MFCC) is considered to be a standard in Audio Speech Recognition (ASR) [28], some approaches have favored spectrograms for audio based diagnosis tasks, arguing that data loss from MFCC could hinder algorithm performance [29]. We opted for the former because of its similarity to how the human cochlea captures sound [30], approximated by a set of non-linear bandpass filters, which could in turn favor the explainability of the model's visualizations.

Audio samples were processed using the MFCC package released by [31], and padded accordingly to generate 6 second long samples, divided into 1 second time steps (see figure 1a for an example sample, and more details on the specific implementation of the MFCC algorithm).

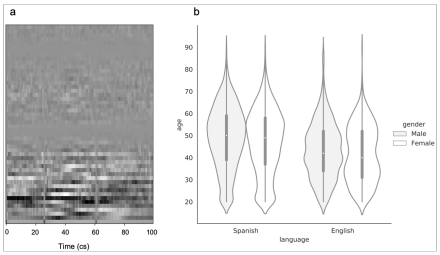

Figure 1: *a. Example 1 second MFCC timestep. All samples were truncated or padded to 6 seconds, and divided into 1 second timesteps. The MFCC generated 50 values per centisecond, with a frequency range of up to 16 kHz, and FFT size of 2048. b. Population distribution of the dataset including age, gender and spoken language. Other collected categorical data included: country, region, symptoms, cough type, COVID-19 diagnosis, diagnosis source and diagnosis date.*

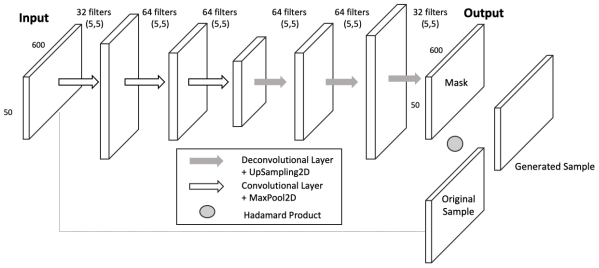

Figure 2: *CAS masking step within the Discovery RNN architecture (trained with Adam [35] and learning rate of 0.001)*

2.2. Classification RNN Models

The first piece of our Discovery RNNs architecture is a recurrent classification model. We opted to train two individual models for the tasks of categorization by *gender* and *spoken language* from our forced coughs dataset. All models were based on the same architecture, which combined a CNN block from residual neural networks (ResNet50) introduced by [32], a single recurrent LSTM layer [33] and two feed-forward dense layers generating the output predictions. Data samples were split in 70% train and 30% test. Train samples were subsequently augmented using a Poisson distribution mask described in equation 1 (where i_x is the input MFCC image, $m(I_x)$ the output mask and $\lambda = 1$), generating an additional 3 recordings per sample.

$$m(i_x) = Pr(\lambda) i_x \tag{1}$$

where

$$Pr(X = k) = \frac{\lambda^k e^{-k}}{k!}$$

2.3. Collaborative Attention Segmentation (CAS) Masking

The second piece of the Discovery RNNs architecture is a Collaborative Attention Segmentation (CAS) model, which takes as input an audio MFCC signal, and outputs a pixel based saliency mask of the same size. In this case, both the encoder and decoder portions of the model are formed by 3 opposing convolutional and deconvolutional blocks, inspired by the U-Net architecture presented in [34]. See figure 2 for a detailed depiction of the CAS architecture.

The Hadamard product (i.e. element-wise multiplication) of this output mask m with the original sample x generates a new MFCC sample of the same size x^*, which can then be fed back into the target classification model τ generating a prediction $\hat{y}^*$.

$$x^* = x \circ m \tag{2}$$

$$\tau(x^*) = \hat{y}^* \tag{3}$$

The loss function for our CAS model, similarly to a traditional categorical cross entropy, ensures that the performance of the target model doesn't degrade between original and generated samples. The goal for this loss function is not to encourage the model to match the exact output value, which can be caused by irrelevant fragments in the sample, but rather to generate a confident class prediction. In order to accomplish this, we use a step function Θ on the output prediction generated from the original MFCC.

$$L(\hat{y^*}, \hat{y}) = -\sum_{j=0}^{C}(\Theta(\hat{y}_j) \cdot \log(\hat{y}_j^*)) \tag{4}$$

We can define this function with input sample x and generated mask m, where τ is the target model.

$$L(x, m) = -\theta\big(\tau(x)\big) \cdot \log\big(\tau(x \circ m)\big) \tag{5}$$

Note that step functions are not differentiable, so in order to train our model with backpropagation, we need to define a continuous approximation function. Within a domain Γ of approximation functions, we can define a function γ as follows, where ϵ is the expected error.

$$\Gamma \supset \gamma(x, m) + \epsilon = L(x, m) \tag{6}$$

We can approximate a binary step function with a modified sigmoid function, defined in equation 7 (where for $k = \inf$, $\sigma(x)$ becomes a step function).

$$\sigma(x) = \frac{1}{1 + e^{-kx}} \tag{7}$$

The cross entropy loss function for an individual sample is then defined as the comparison between predictions of the target model for both original and generated MFCC samples (see equation 8).

$$\gamma_1(x, m) = -\sigma\big(\tau(x)\big) \cdot \log\big(\tau(x \circ m)\big) \tag{8}$$

To encourage the model to deactivate as many pixels as possible, we define a second loss function which computes the average pixel value of the mask, or density, so the model doesn't default to an empty mask, which would in turn generate an output MFCC sample indistinguishable from the original (see equation 9, where W and H are the width and height of the mask M).

$$\gamma_2(m) = \frac{1}{W \cdot H}\sum_{i=0}^{W}\sum_{j=0}^{H}(m_{ij}) \tag{9}$$

The final loss function, is defined as the weighted sum of both of these, where α is selected as a hyperparameter.

$$\gamma(x, m) = \gamma_1(x, m) + \alpha\gamma_2(m) \tag{10}$$

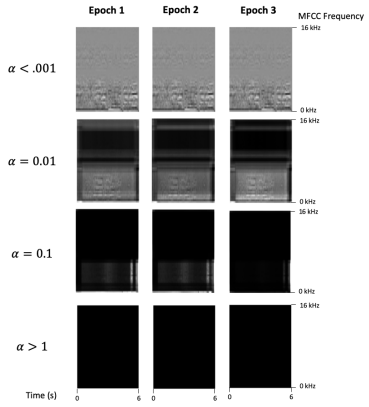

Figure 3: *Generated images from CAS model trained on spoken language classification, for four α values over 3 training epochs. α values above 0.1 stabilize on a full mask (generating an empty MFCC), while values below 0.001 stabilize on an empty mask (generating an MFCC indistinguishable from the original).*

2.4. Expert Cognitive Bias Loop

If we were to optimize our model with the ideal step function Θ, we would need an α small enough to make sure that no well classified sample became miss-classified. In fact, an infinitesimally small α would work. Since our computational methods require continuous functions, we need an α larger than infinitesimally zero to offset errors of approximation. Hence, we can state for σ:

$$\exists\alpha > 0 : \epsilon < \alpha \tag{11}$$

Following this, it appears that α is not a hyperparameter to optimize, but rather a setting to navigate the hyperspace Γ of approximated functions, with the goal of identifying different local minima. How these local minima are identified and interpreted is what allows for expert input to guide the training and optimization of our Discovery RNNs architecture.

3. Results

The RNN model reached 67.6% and 78.6% balanced accuracies for the tasks of spoken language and gender classification respectively. Figure 3 depicts generated MFCC samples for several α values for the target model of spoken language classification. Note that generated images with α values above 1 stabilize on a null image, as the density loss (equation 9) overshadows the effects of the crossentropy loss (equation 8). The opposite is true for α values below 0.01. Within this bracket, higher values ($\alpha = 0.1$) generate strong low frequency and temporal segmentation masks in the first 2 epochs of training, finally stabilizing to a null image once again. Lower values ($\alpha = 0.01$) stabilize on a low frequency segmentation mask over time.

Figure 4 shows four generated samples from the test dataset with their average activation over time, from the CAS model for the task of spoken language classification ($\alpha = 0.1$). Results show a bias towards the end of the recording, sample 4 showcases how this effect is increased when the sample is padded, suggesting this as a plausible explanation. Average activation for samples 1, 2 and 4 show the model pays attention to the beginning of a cough event, and is able to identify multiple independent cough events within the same sample, as portrayed

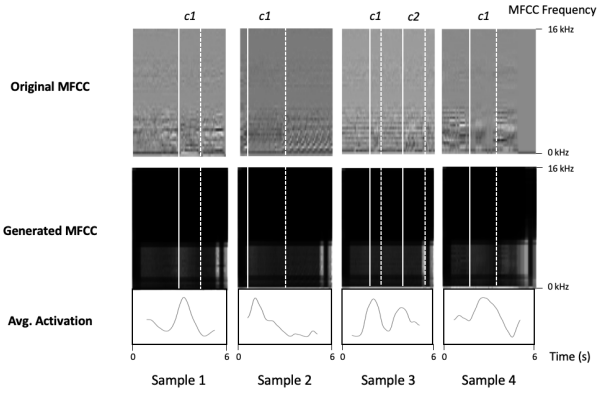

Figure 4: *Four generated and reference MFCCs from the test dataset, for the task of spoken language classification ($\alpha = 0.1$ and epoch 1), where solid and dashed white lines denote the beginning and end of a cough event respectively. Generated masks show a negative bias towards the beginning of all recordings, and a positive bias towards the end. Average activation per sample was computed for timesteps in-between these biases for the sake of visualization. For all samples, average activation reliably identifies the beginning of a cough event, independently of whether the sample is padded or not.*

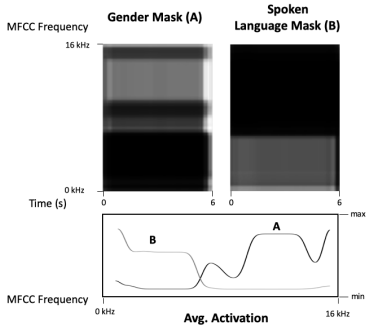

Figure 5: *Generated masks from CAS model trained on gender and spoken language respectively ($\alpha = 0.01$, epoch 10), and average activation over the frequency domain. Gender masks show a clear bias towards higher frequencies, while spoken language masks show a bias towards low frequencies.*

in sample 3.

For the task of gender classification, generated images did not identify cough events reliably. Figure 5 shows the masks from a single input sample, for both CAS models trained for the task of gender and spoken language classification respectively. Notice how masks for the gender model segment high frequencies, as opposed to masks for the spoken language classification task, which segment low frequencies. This was the case for all reviewed generated samples for the gender classification task. Gender classification samples show a similar bias towards the end of the sample similarly to the spoken language classification visualization with $\alpha = 0.1$.

4. Discussion

Balanced accuracies for both RNN classification tasks reveal for the first time that forced coughs can be used to extract information beyond the simple characteristics of the symptom. These results suggest that a holistic approach to cough analysis could be taken, to evaluate what features could be used not only for biological and cultural identification purposes, as portrayed here, but in medical diagnosis and screening. Using forced coughs specifically could enable the study of asymptomatic patients in the context of the COVID-19 pandemic.

Specifically for the spoken language task, visualizations from the Discovery RNN models (both for $\alpha = 0.1$ and $\alpha = 0.01$) confirm that these predictions are based on the cough event specifically, as well as identifying a possible source of bias on the last portion of the cough samples caused by the preliminary padding of short samples. Generated images for high α values identified the beginning of the cough event, confirming that predictions were not solely based on a dataset bias. The model proved to be robust in samples with multiple independent cough events, accurately identifying multiple cough events in a single sample. Note that this segmentation was only based on model saliency, future work could evaluate this method in other audio domains and computer vision to validate this architecture for unsupervised object or audio event segmentation.

For the task of gender classification, generated masks confirmed a bias towards padded images. In this case, low α values segmented high frequencies as the source of model prediction, while failing to identify individual cough events, suggesting that model predictions are in this case based on environmental or non-cough related cues. This result is surprising, given that the gender classification task is notably easier for a human to perform than the spoken language task. Arguably, identifying and cleaning the dataset of other biases could affect these results.

Overall, Discovery RNN generated samples provide a first stepping stone towards visualizing RNN saliency in audio samples, potentially enabling the transfer of untapped neural network knowledge for physicians. The fact that these algorithms are based on MFCCs, which are somewhat related to how the human ear captures sound, as opposed to temporal spectrograms, suggests that physicians could be trained to perform similar tasks. Future work could explore this concept by setting up a training pipeline for physicians to learn to classify forced coughs based on the areas of interest defined by the Discovery RNN models. Another interesting application of Discovery RNNs could be the identification of biased models and datasets, such as the one exposed in this paper. More work is warranted in this direction to confirm whether these results are transferable to other audio based domains.

5. Conclusions

RNN visualization and explainability stands as a major challenge towards implementing audio analysis based diagnosis and screening in medical environments. Developing algorithms to better understand how predictions are made, and potentially transferring this knowledge to physicians so they can take better decisions, could prove fundamental in this regard. In this paper we present a novel approach to visualize model saliency in a forced cough recordings dataset for COVID-19 in the tasks of spoken language and gender classification. We were able for the first time to demonstrate that biological and cultural information such as gender and spoken language respectively can be inferred from forced cough analysis. Furthermore, our approach allowed us to identify a bias in our dataset, while robustly segmenting cough events in an unsupervised manner. Future work in this direction could include evaluating Discovery RNNs for the tasks of COVID-19 screening and diagnosis, and the role of physicians in guiding the training of these algorithms.

6. References

[1] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," *arXiv preprint arXiv:1611.03530*, 2016.

[2] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*, 2016, pp. 173–182.

[3] Y.-A. Chung, C.-C. Wu, C.-H. Shen, H.-Y. Lee, and L.-S. Lee, "Audio word2vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder," *arXiv preprint arXiv:1603.00982*, 2016.

[4] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling, "Visualizing deep neural network decisions: Prediction difference analysis," *arXiv preprint arXiv:1702.04595*, 2017.

[5] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[6] T. Davenport and R. Kalakota, "The potential for artificial intelligence in healthcare," *Future healthcare journal*, vol. 6, no. 2, p. 94, 2019.

[7] P. Tschandl, C. Rosendahl, B. N. Akay, G. Argenziano, A. Blum, R. P. Braun, H. Cabo, J.-Y. Gourhant, J. Kreusch, A. Lallas *et al.*, "Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural networks," *JAMA dermatology*, vol. 155, no. 1, pp. 58–65, 2019.

[8] P. Teare, M. Fishman, O. Benzaquen, E. Toledano, and E. Elnekave, "Malignancy detection on mammography using dual deep convolutional neural networks and genetically discovered false color input enhancement," *Journal of digital imaging*, vol. 30, no. 4, pp. 499–505, 2017.

[9] G. Bedi, F. Carrillo, G. A. Cecchi, D. F. Slezak, M. Sigman, N. B. Mota, S. Ribeiro, D. C. Javitt, M. Copelli, and C. M. Corcoran, "Automated analysis of free speech predicts psychosis onset in high-risk youths," *npj Schizophrenia*, vol. 1, p. 15030, 2015.

[10] T. Al Hanai, M. M. Ghassemi, and J. R. Glass, "Detecting depression with audio/text sequence modeling of interviews." in *Interspeech*, vol. 2522, 2018, pp. 1716–1720.

[11] T. Alhanai, R. Au, and J. Glass, "Spoken language biomarkers for detecting cognitive impairment," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 409–416.

[12] R. Poplin, A. V. Varadarajan, K. Blumer, Y. Liu, M. V. McConnell, G. S. Corrado, L. Peng, and D. R. Webster, "Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning," *Nature Biomedical Engineering*, vol. 2, no. 3, p. 158, 2018.

[13] A. Morice *et al.*, "The diagnosis and management of chronic cough," *European Respiratory Journal*, vol. 24, no. 3, pp. 481–492, 2004.

[14] R. Palaniappan, K. Sundaraj, and N. U. Ahamed, "Machine learning in lung sound analysis: a systematic review," *Biocybernetics and Biomedical Engineering*, vol. 33, no. 3, pp. 129–135, 2013.

[15] G. J. Criner, J. Bourbeau, R. L. Diekemper, D. R. Ouellette, D. Goodridge, P. Hernandez, K. Curren, M. S. Balter, M. Bhutani, P. G. Camp *et al.*, "Executive summary: prevention of acute exacerbation of copd: American college of chest physicians and canadian thoracic society guideline," *Chest*, vol. 147, no. 4, pp. 883–893, 2015.

[16] B. Subirana, F. Hueto, P. Rajasekaran, J. Laguarta, S. Puig, J. Malvehy, O. Mitja, A. Trilla, C. I. Moreno, J. F. M. Valle *et al.*, "Hi sigma, do i have the coronavirus?: Call for a new artificial intelligence approach to support health care professionals dealing with the covid-19 pandemic," *arXiv preprint arXiv:2004.06510*, 2020.

[17] D. Cucinotta and M. Vanelli, "Who declares covid-19 a pandemic." *Acta bio-medica: Atenei Parmensis*, vol. 91, no. 1, pp. 157–160, 2020.

[18] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," *University of Montreal*, vol. 1341, no. 3, p. 1, 2009.

[19] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, 2015.

[20] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune, "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks," in *Advances in neural information processing systems*, 2016, pp. 3387–3395.

[21] A. Mahendran and A. Vedaldi, "Visualizing deep convolutional neural networks using natural pre-images," *International Journal of Computer Vision*, vol. 120, no. 3, pp. 233–255, 2016.

[22] L. Bazzani, A. Bergamo, D. Anguelov, and L. Torresani, "Self-taught object localization with deep networks," in *2016 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2016, pp. 1–9.

[23] A. Nguyen, J. Yosinski, and J. Clune, "Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks," *arXiv preprint arXiv:1602.03616*, 2016.

[24] Z. Tang, Y. Shi, D. Wang, Y. Feng, and S. Zhang, "Visualization analysis for recurrent networks," *Tech. Rep.*, 2016.

[25] A. Karpathy, J. Johnson, and L. Fei-Fei, "Visualizing and understanding recurrent networks," *arXiv preprint arXiv:1506.02078*, 2015.

[26] G. Keren, M. Schmitt, T. Kehrenberg, and B. Schuller, "Weakly supervised one-shot detection with attention siamese networks," *stat*, vol. 1050, p. 12, 2018.

[27] T.-W. Su, J.-Y. Liu, and Y.-H. Yang, "Weakly-supervised audio event detection using event-specific gaussian filters and fully convolutional networks," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 791–795.

[28] W. Han, C.-F. Chan, C.-S. Choy, and K.-P. Pun, "An efficient mfcc extraction method in speech recognition," in *2006 IEEE international symposium on circuits and systems*. IEEE, 2006, pp. 4–pp.

[29] H. Lee, P. Pham, Y. Largman, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Advances in neural information processing systems*, 2009, pp. 1096–1104.

[30] J. H. McDermott and E. P. Simoncelli, "Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis," *Neuron*, vol. 71, no. 5, pp. 926–940, 2011.

[31] J. Lyons, "Python speech features," *Available at github. com/jameslyons/python_speech_features*, 2017.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[33] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[34] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.