GAN-based Generation and Automatic Selection of Explanations for Neural Networks

Saumitra Mishra*[1], Daniel Stoller*, Emmanouil Benetos*#, Bob L. Sturm† & Simon Dixon*

*School of EECS, Queen Mary University of London, UK
#The Alan Turing Institute, UK
†Speech, Music and Hearing, KTH Royal Institute of Technology, Sweden

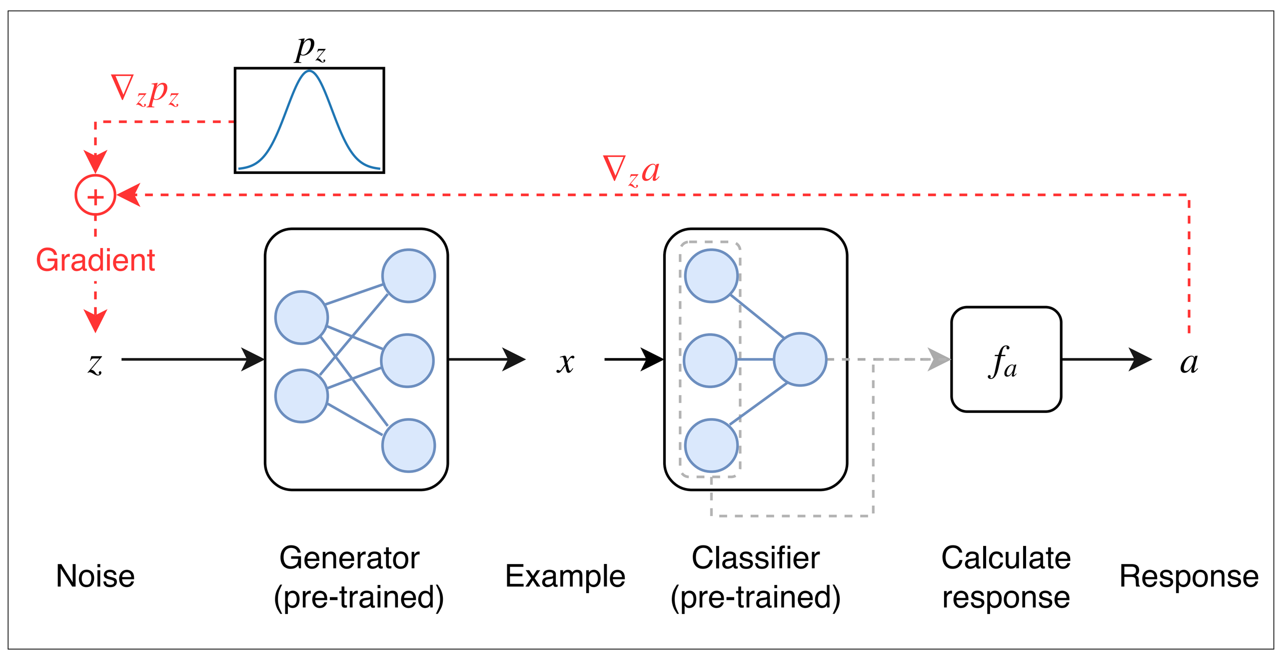
[1]saumitra.mishra@qmul.ac.uk
https://sites.google.com/site/saumitramishrac4dm/

Overview

- We use a Generative Adversarial Network (GAN) to generate realistic explanations using Activation-Maximisation (AM) [1]. We validate our method on a vocal classifier, showing it can retrieve the concept of singing voice presence encoded in the output layer neuron.
- We also propose Fréchet Inception Distance (FID) [2] as a quantitative measure for estimating the interpretability of a set of generated examples. We demonstrate the effectiveness of FID in automatically evaluating the interpretability of a set of generated explanations.

GAN-based Activation Maximisation

- AM synthesizes examples (e.g., images) that maximally activate different components (neurons, layers) of a deep neural network (DNN).
- To generate visually interpretable examples, AM uses regularisers that restrict the search space to realistic examples.



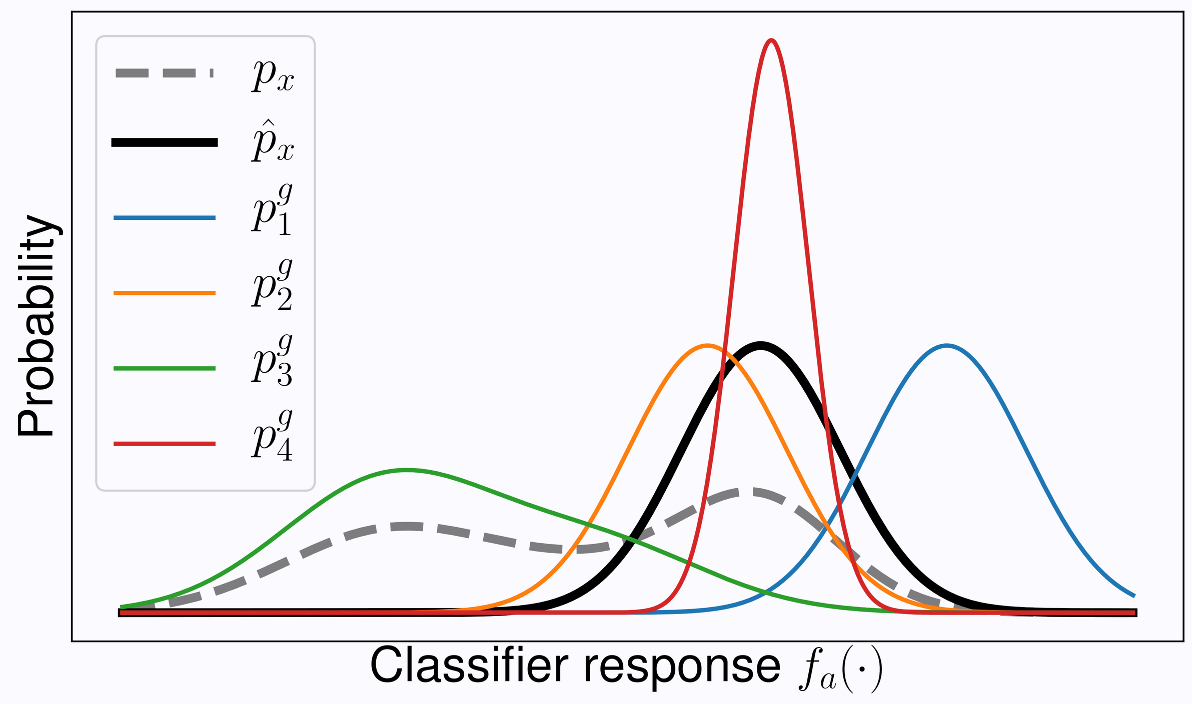
- Our method uses a GAN as a regulariser that imposes a strong prior. The method optimises

$$\hat{z} = \arg\max_{z} f_a(f_n(f_g(z))) + \lambda \log p_z(z). \quad (1)$$

  - $f_g : \mathbb{R}^n \to \mathbb{R}^d$ represents a generator that maps a noise vector $z \in \mathbb{R}^n$ drawn from a known noise distribution p_z to a generated example $x \in \mathbb{R}^d$
  - $f_n(x) \in \mathbb{R}^M$ represents activations of *all* M neurons in a neural network classifier f_c
  - $f_a : \mathbb{R}^M \to \mathbb{R}$ represents classifier response
  - $\lambda \geq 0$ controls the trade-off between AM and the realism of the generated examples

Quantitative Selection of Explanations

- Present methods for selecting optimal AM hyper-parameters rely on visual interpretability of generated examples [3], but this is time-intensive and does not scale well.
- We propose **Fréchet Inception Distance (FID)** as a metric for efficiently evaluating the interpretability of a set of generated explanations.



- We posit that good interpretability requires the generated examples to have a similar distribution of classifier responses $f_a(\cdot)$ as the N samples with the highest response from the dataset ($\hat{p}_x$).
- We select hyper-parameters that minimise FID between the dataset and the generated response distributions.

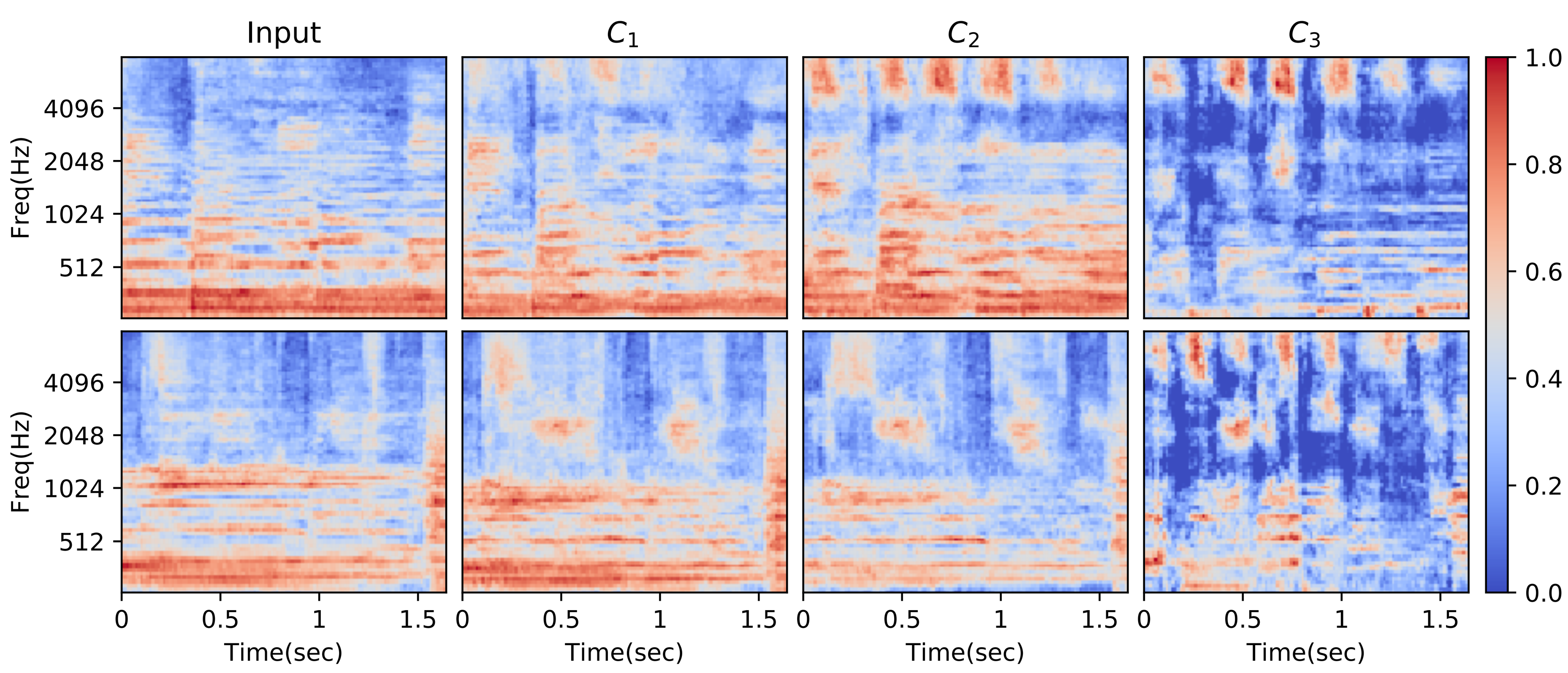
Experiments

- **Classifier**- State-of-the-art audio classification model that classifies a time-frequency representation of an audio excerpt to the "vocal" or "non-vocal" class. The model is an 8-layer deep variant of VGG-Net with a single sigmoidal neuron in the output layer.
- **GAN training**
  - Generator noise distribution $p_z(z) = \mathcal{N}(z|\mathbf{0}_n; \mathbf{I}_n)$, where $n = 128$
  - Generator and discriminator architectures are variants of DCGAN
- **AM optimisation**
  - Learning rate $l_r \in \{0.1, 0.01, 0.001\}$, prior weight $\lambda \in \{0.1, 0.01, 0.001\}$, number of iterations $N_t \in \{100, 500, 1000\}$, number of examples per hyper-parameter configuration $N = 50$

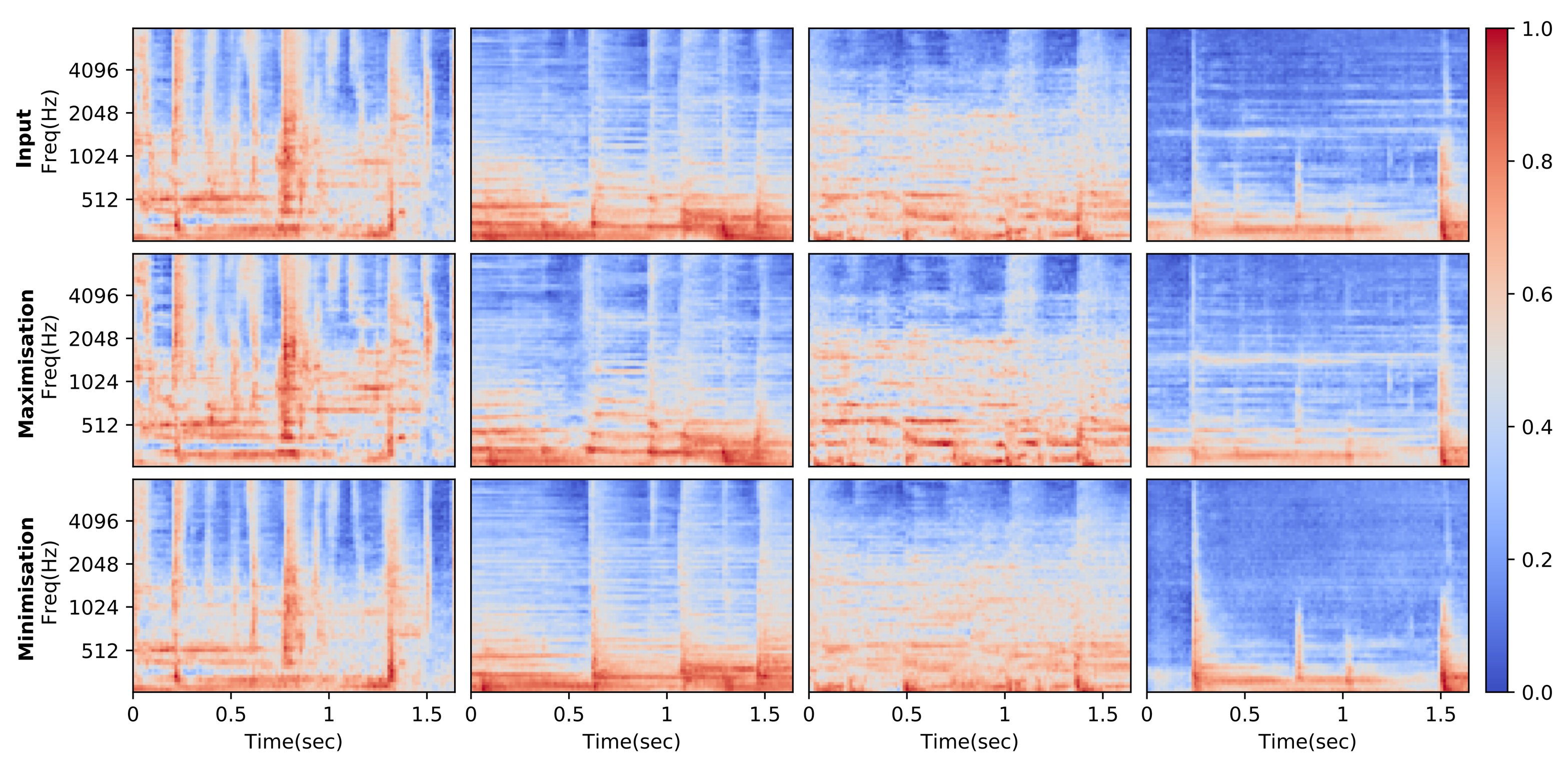
[1] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox and J. Clune. Synthesizing the Preferred Inputs for Neurons in Neural Networks via Deep Generator Networks. In *Proc. NeurIPS*, 2016.

[2] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler and S. Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Proc. NeurIPS*, 2017.

Results



(a) **Hyper-parameter optimisation** - FID successfully quantifies the interpretability of generated examples. C_1, C_2 and C_3 represent the best, median and worst hyper-parameter configurations (from FID), respectively.



(b) **Qualitative analysis of explanations** - We use AM to maximise and minimise the response of the output layer neuron resulting in vocal and non-vocal examples, respectively.