EXPLORING EXPLAINABILITY FOR TIME SERIES REPRESENTATIONS USING THE RELAX METHOD

David Elias Hammershøi (s204654) , Nusret Emirhan Salli (s204658), Katja Valérie Bonvin (s233494)

Supervisors: Thea Brüsch & Teresa Karen Scheidt

ABSTRACT

Interpreting the decisions of deep neural networks in the context of time series data presents unique challenges. This paper explores the interpretability of a model using the RELAX method. The model was trained in classifying animal sounds based on Mel spectrograms. Originally designed for images, the RELAX method employs unstructured random masking to highlight model decision boundaries and describe the localization and faithfulness of the model. The results from this paper indicate that unstructured random masking lacks discriminatory power across time series data, while structured masking reveals nuanced attention patterns. Pixel flipping experiments validate the importance assigned by RELAX but expose challenges in direct transferability to spectrograms. The study underscores the sensitivity of explainability methods to data characteristics, emphasizing the need for tailored approaches to time series representations. Future work includes collaboration with domain experts, refining methods for spectrograms, and comparative analyses with alternative explainability methods. This research contributes insights into the challenges and potential advancements in enhancing the transparency and reliability of neural network decision interpretations, particularly in the realm of time series data.

***Index Terms*—** Explainable AI, Mel spectrograms, Time series data, RELAX, Uncertainty

1. INTRODUCTION

Explainable AI (XAI) is crucial for understanding weaknesses and biases, especially in time series representations where patterns may be less apparent compared to other data types such as images. XAI not only fosters trust but also contributes to verifying decisions, crucial in both supportive and autonomous roles. Recent AI advancements in uncovering hidden patterns in scientific data highlight the potential for groundbreaking discoveries. Representation Learning Explainability (RELAX) is a method for explaining representations, equipped with uncertainty quantification, which has proven to work quite well in computer vision tasks. Hence this paper intends to explore the use of the RELAX method for explainability for times series representations instead.

2. METHODS

2.1. Data Prepocessing & Selection

The objective was to use easily interpretable and repeatable data, such that results from an explanation would be as easy to interpret as possible. With this in mind, a classification task focused on distinguishing between Mel spectrograms of three distinct animal categories: dogs, cats, and birds was chosen.
A dataset of 200 Mel spectrograms was created from 975 ms audio inputs by taking the raw pulse code modulation from the WAV file and converting them into floating-point values scaled between -1.0 and +1.0 [1]. The data was then resampled to a standardized rate of 16,000 samples per second. Following resampling, the data underwent segmentation into overlapping windows, with each window systematically subjected to a Hamming Window for further processing. The Power Spectrum was derived through a Fast Fourier Transformation, where frequencies above and below specific thresholds systematically were eliminated to refine the processed data. Mel Frequency Filter Banks were subsequently applied, introducing another layer of transformation. Finally, a logarithmic function was applied to all values and a Mel spectrogram of shape (96, 64) was created. This corresponds to sound clips with a duration of 0.96 seconds, or approximately 1 second.

2.2. Pre-trained model: VGGish

The model used for feature extracting is a modified version of the VGG model, VGGish [2] [3]. VGG stands for Visual Geometry Group, and it is a standard deep CNN architecture with multiple layers, whose architecture can be seen in Figure 1.
VGGish is a pre-trained Convolutional Neural Network developed by Google, trained on an extensive dataset comprised of audio clips extracted from 100 million YouTube videos, annotated with 3,000 distinct labels with cross-entropy loss as its loss function. The model has a sequence of convolution and activation layers, followed by a max pooling layer where this neural network is structured with a total of 17 layers. Following the pre-training the last convolutional layer was removed while the softmax was removed as well as the

fully connected layer being changed to be 128-wide. Thus the output is a 128-wide embedding layer that can be used in transfer learning scenarios, such as animal sound classification.

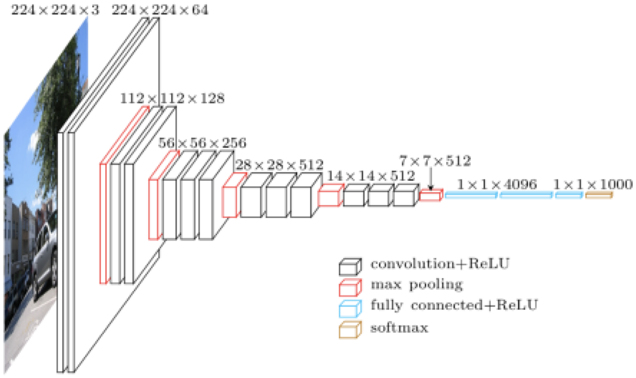

Fig. 1. VGG Neural Network Architecture. [4]

To validate the correct implementation of the model and assess its ability to extract meaningful features from animal sounds, a logistic regression classifier was trained on top of the embedding layer. This classification task aimed to distinguish between dogs, birds, and cats in the latent space. With the dataset consisting of 200 animal sound spectrograms, a classification accuracy of 97% was achieved on the test dataset which had a total of 40 observations. This high accuracy suggests that the model was able to extract important features from the classes for the classifier to easily distinguish between them. The observations in the latent space can be seen in Figure 2, which clearly shows how the different classes are generally speaking represented uniquely in the latent space.

2.3. RELAX: Representation Learning Explainability

The RELAX method employs an occlusion-based approach by masking sections of the Mel spectrogram. The key metric used for calculating saliency is the cosine similarity between the unmasked and masked representations [5]. This similarity measure is pivotal in distinguishing between non-informative and informative regions.

The explanations are computed by sampling N masks and computing the sample mean referred to as saliency. The uncertainty of each pixel is also calculated by calculating how much the similarity changes from mask to mask for specific pixels where large changes from mask to mask would yield a high uncertainty value. Formally speaking, this can be understood as the variance of the saliency. The cosine similarity is defined as:

$$s(\mathbf{h}, \bar{\mathbf{h}}) = \frac{\langle \mathbf{h}, \bar{\mathbf{h}} \rangle}{|\mathbf{h}||\bar{\mathbf{h}}|}.$$

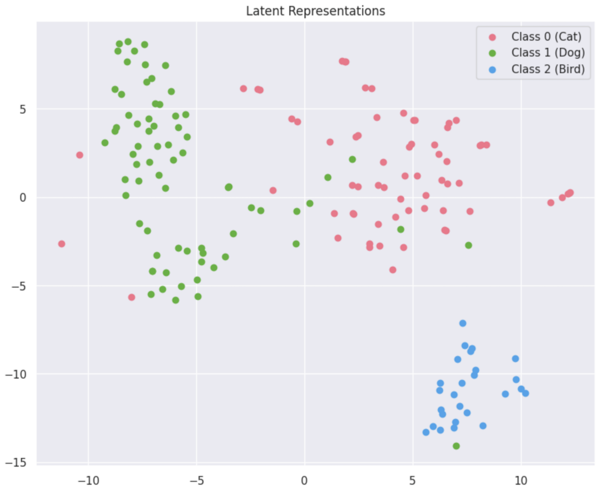

Fig. 2. Visualization of the latent space representation of the samples computed using T-distributed Stochastic Neighbor Embedding.

Here, $\mathbf{h}$ denotes the unmasked representation, and $\bar{\mathbf{h}}$ represents the masked version. The similarity is expected to be high when non-informative parts are masked out and low when informative elements are masked out. The explanations are computed by sampling N masks and computing the sample mean, used as a way to localize the importance of a spectrogram

$$\bar{R}_{ij} = \frac{1}{N} \sum_{n=1}^{N} s(\mathbf{h}, \bar{\mathbf{h}}_n) M_{ij}(n)$$

where $\bar{\mathbf{h}}_n$ is the representation of the image masked with mask n, and $M_{ij}(n)$ the value of element (i, j) for mask n. The uncertainty of the RELAX-score for pixel (i, j):

$$\bar{U}_{ij} = \frac{1}{N} \sum_{n=1}^{N} (s(\mathbf{h}, \bar{\mathbf{h}}) - \bar{R}_{ij})^2 M_{ij}(n).$$

This formulation offers valuable insights into the uncertainty linked with the contribution of individual pixels to the overarching saliency score, as discussed in [5]. These insights provide a deeper understanding of the faithfulness of the saliency measurement.

The reason for selecting the cosine similarity to measure the similarities between the masked and unmasked spectrogram has also been considered. The main point is that cosine similarity is scale-invariant since it calculates the angle between the 2 points compared to calculating magnitude information (which for instance the Euclidean distance does), which makes it more robust in higher dimensionality [5].

2.4. Masking Strategies

As the RELAX method operates as an occlusion technique, it is therefore a necessity to implement masking strategies. In alignment with the original RELAX paper's approach to images, an unstructured random masking method is applied to the Mel spectrograms. In addition to this unstructured random masking, a structured masking approach is implemented, targeting specific portions in terms of time, frequency, or a combination of both as shown in Figure 3. The replacement value of the masked area was set to the mean values of the given spectrogram. This decision was made to ensure that the replacement value, unlike zero as in an image, would convey meaningful information, considering alternatives such as minimum or maximum values.

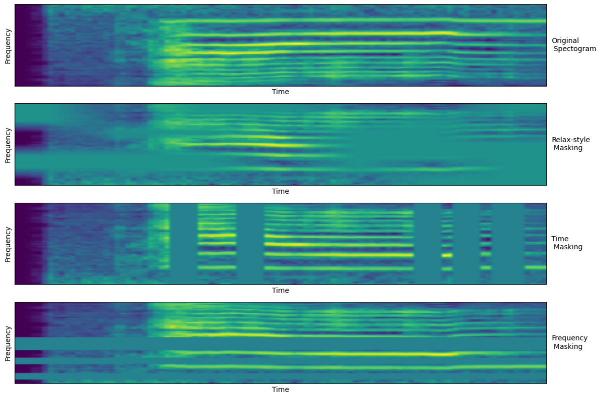

Fig. 3. Various masking strategies employed in the RELAX method, with the mean as the replacement value.

3. RESULTS

The RELAX method was employed to enhance representation learning explainability, utilizing different masking strategies to uncover the model's decision-making processes. Saliency was calculated based on occlusion, where masking parts of the Mel spectrogram allowed the computation of cosine similarity between the unmasked and masked representations. The explanations were generated by sampling masks and computing the saliency. Additionally, the uncertainty of each pixel's contribution to the overall saliency score was quantified by measuring the change in similarity between masks for specific pixels. This provided insights into the uncertainty associated with each pixel's interpretability, enhancing the overall transparency of the model's decision-making process. Furthermore, to verify our findings further pixel-flipping has been used to investigate its faithfulness [6].

3.1. Unstructured Random Masking

Unstructured random masking, applied uniformly across all Mel spectrograms, as was done in the RELAX paper, failed to reveal clear explainable patterns shown in Figure 4. The results indicated a consistency in the masking impact across all samples, with almost no variation in the saliency scores among the 3 classes.

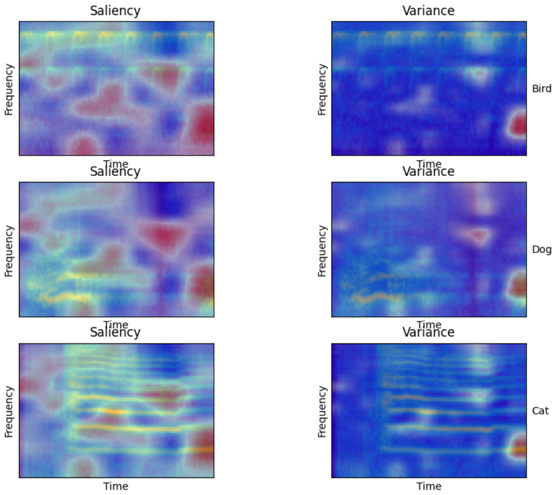

Fig. 4. Saliency and variance score, where red indicates high importance and blue indicates low importance, obtained using random masking.

This lack of differentiation suggested that the unstructured random masking approach did not effectively highlight specific features or contribute to the interpretability of the model's decision-making when differentiating between the classes.

3.2. Structured Random Masking

In contrast, the structured masking strategy demonstrated a more clear pattern. This approach, targeting specific portions of the spectrograms in terms of time, frequency, or a combination of both, revealed distinct patterns in the model's attention. Particularly noteworthy was the emphasis on silent parts of the spectrograms as seen in Figure 5.

The structured masking approach effectively highlighted the model's attention to what humans might perceive as unintuitive, yet it revealed a consistently clear pattern for the three distinct animal sounds, with a focus on the silent parts of the spectrogram.

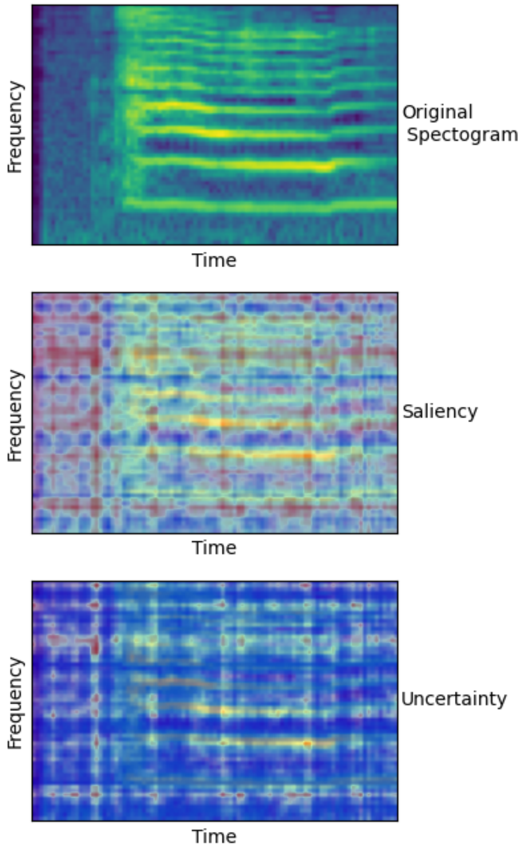

Fig. 5. Example of results from time- and frequency-based masking for the sound of a cat showing the tendency of the model to focus on the more silent parts of the spectogram.

3.3. Pixel Flipping

Flipping the identified important pixels, as determined by the RELAX method, is a strategic approach to further investigate the robustness and interpretability of the model's decisions. By selectively altering these key pixels and observing the corresponding changes in the model's output, we are able to gain deeper insights into the specific features that drive the model's predictions. This process essentially involves validating the importance assigned to specific pixels by the RELAX method and understanding their impact on the overall decision boundaries. Such pixel-flipping experiments contribute to the overall assessment of the explainability and reliability of the model's insights generated by the RELAX method.

The result from pixel flipping can be seen in figure 6. It is evident based on the figures that the method does find important structures in the spectrograms, as when these are set to the mean or minimum, we see a steep decrease in the Cosine similarity.

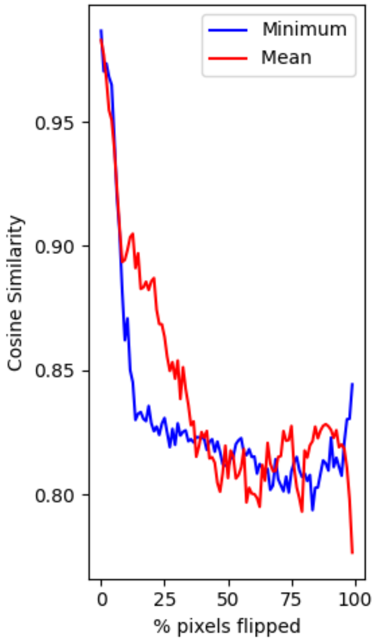

Fig. 6. The figure illustrates Pixel Flipping, a technique wherein pixels are selectively removed based on their importance calculated from the RELAX method using time-frequency masking. Subsequently, the similarity between the original image and the altered version is computed: one with pixel replacements using the mean value, and the other utilizing the minimum value from the Mel spectrogram.

4. DISCUSSION

The reliance on human judgment introduces subjectivity in determining important aspects, a common challenge in explainability. Furthermore, patterns that the Deep neural networks are able to extract which humans might not be able to comprehend can lead to incompatibilities between the model and the individual that needs to get insights into the model's decision process.

In a broader context, if the applied masking strategies demonstrate negligible effects on the saliency map or yield nothing but noise, as the unstructured random masking did, a logical reason for this would be that the chosen method for quantifying explainability may not be applicable. This is a valid concern since the RELAX method is based on images and thus might not even be applicable for time series representation even if the time series data have been transformed into Mel spectrograms. However, the structured random masking showed a clear and consistent pattern, but also had high uncertainty in places where there was high saliency. This indicates that the findings are uncertain, thus having a low faithfulness and in turn usefulness. Hence, our experimentation with time- and frequency masking and random masking has revealed distinctly disparate outputs. The divergence in results between these two masking strategies highlights the explainability method's sensitivity to the choice of masking technique. Therefore it is also important to investigate and verify the findings, which in this case was found based on pixel-flipping. This evaluation metric would be a way of testing the method's faithfulness. The results from figure 6 show that by "flipping" the values to the mean value the cosine similarity decreases quite rapidly until it hits a "plateau" between 80-85% and keeps oscillating. A plausible explanation is that information might not be "removed" since the value would be changed to mean and thus still retain some information. Similar results were obtained even if the pixels were flipped to the minimum value which is the usual procedure for images and once again highlights that methods (for instance RELAX and pixel flipping) might not be directly applicable to spectrograms. As a sanity check the similarity between a spectrogram with all values set to the mean and an animal sound was also computed, and found to be approximately 75%.

An optimal method for measuring the faithfulness of RELAX's output would involve constructing a pixel-flipping curve based on metrics such as accuracy or mean error. This would allow examination of whether or not the pixels identified as important by the method correspond to an impact on the task of classification. Unfortunately, this was not executed due to the computational expense of both RELAX and pixel flipping. Running these methods on an extensive dataset would be necessary to generate reliable results.

4.1. Future work

Engaging with experts in animal sounds can provide valuable insights to fine-tune interpretative decisions and mitigate potential biases, thereby fortifying the robustness of our explainability approach. Conducting experiments by removing parts of sounds based on importance for expert evaluation offers a unique opportunity. While this will alter the original animal sounds, it serves as a valuable test to assess the experts' ability to distinguish and identify the sounds under varied conditions.

An alternative method to explore the underlying patterns within the spectrogram could be examining outliers in the latent representation. Notably, one of the dog sounds appears highly similar to all bird sounds. A more in-depth investigation may reveal clear patterns in the spectrogram, providing a way to examine what is important based on the latent space.

As mentioned earlier, the approach to use the RELAX method and other methods used on images might be flawed in its current state, Modifying the RELAX method to better suit spectrograms might give more useful and consistent results to gain insight on. Furthermore, the RELAX method is an occlusion-based method but other methods have also been developed to gain insight into the model's decision. Methods such as Integrated gradients and Layer-wise Relevance Propagation can be implemented to effectively compare different methods [6]. This could be used to investigate whether or not different methods give out drastically different results.

5. CONCLUSION

This paper addresses interpretability challenges in deep neural networks applied to time series data, specifically the classification of animal sounds represented as Mel spectrograms. Our exploration of the RELAX method, initially designed for images, yields nuanced findings. Unstructured random masking proves inadequate for explaining the distinguishing between animal classes, underscoring the limitations of direct transferability to time series data. In contrast, structured masking offers clearer insights into the model's attention patterns. Pixel flipping experiments validate RELAX's ability to identify important structures based on similarity, emphasizing the need for careful interpretation and adaptation of methods across diverse data types. The divergence in results between time-frequency masking and random masking highlights the necessity for tailored approaches in XAI for time series data. While the model demonstrated good accuracy, showing its capability to extract meaningful patterns for classification, applying the RELAX method revealed no significant explanatory patterns, at least not perceptible to our analysis. This could be attributed to either the inherent uninterpretability of underlying patterns or the potential limitation of the RELAX method when applied to spectrogram representations. Given the subjective nature of interpretability, collaboration with domain experts is crucial. Future research should refine methods tailored for spectrograms and conduct comparative analyses with alternative XAI approaches. This study contributes valuable insights into interpreting neural network decisions in time series data, paving the way for advancements in transparent and reliable model explanations.

6. REFERENCES

[1] YashNita, "Animal-sound-dataset," https://github.com/YashNita, 2018.

[2] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin W. Wilson, "CNN architectures for large-scale audio classification," *CoRR*, vol. abs/1609.09430, 2016.

[3] Harri Taylor, "torchvggish," `https://github.com/harritaylor/torchvggish/tree/master`, 2021.

[4] VGG, "Vgg very deep convolutional networks," 2023.

[5] Kristoffer K. Wickstrøm, Daniel J. Trosten, Sigurd Løkse, Ahcène Boubekki, Karl Øyvind Mikalsen,

Michael C. Kampffmeyer, and Robert Jenssen, “Relax: Representation learning explainability,” *International Journal of Computer Vision*, , no. 131, pp. 1584–1610, 2023.

[6] Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Cristopher J. Anders, and Klaus-Robert Müller, “Explaining deep neural networks and beyond: A review of methods and applications,” *IEEE*, vol. abs/1609.09430, 2021.

7. APPENDIX

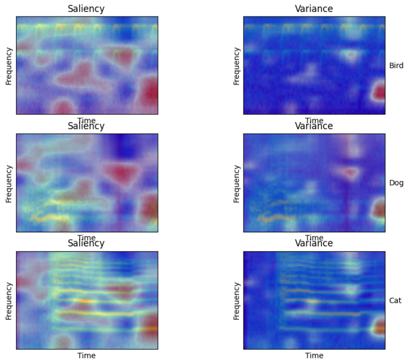

Fig. 7. Saliency and variance score, where red indicates high importance and blue indicates low importance, obtained using random masking.

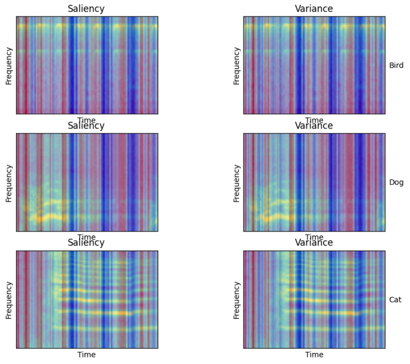

Fig. 8. Saliency and variance score, where red indicates high importance and blue indicates low importance, obtained using time masking.

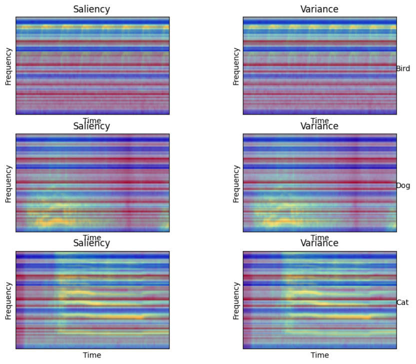

Fig. 9. Saliency and variance score, where red indicates high importance and blue indicates low importance, obtained using frequency masking.

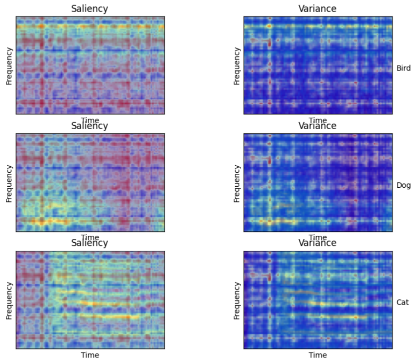

Fig. 10. Saliency and variance score, where red indicates high importance and blue indicates low importance, obtained using time and frequency masking.