Exploring Explainability for Time Series Representations

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

Deep Learning 2023 - 02456, Technical University of Denmark

Motivation

- Explainable AI (XAI) is crucial for understanding weaknesses and biases, especially in time series representations where patterns may be less apparent
- Contribution in the verification of decisions
- RELAX is a method for explaining representations, equipped with uncertainty quantification, which has proven to work quite well in computer vision tasks [5]

Furthermore to verify the results from an XAI model several evaluation metrics should be considered:

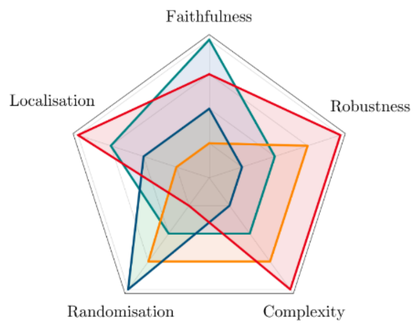


Figure 1: Different evaluation metrics [2]

Key Points

- The deep CNN model VGGish has been employed, which is based on the standard deep CNN Visual Geometry Group (VGG) model [3][1]
- A logistic classifier has been made to validate the VGGish' ability to classify animal sounds based on spectrogram features
- Different masking strategies have been implemented to sufficiently determine how the model predicts
- We compared the spectrograms with the RELAX explanations

VGGish Neural Network Architecture

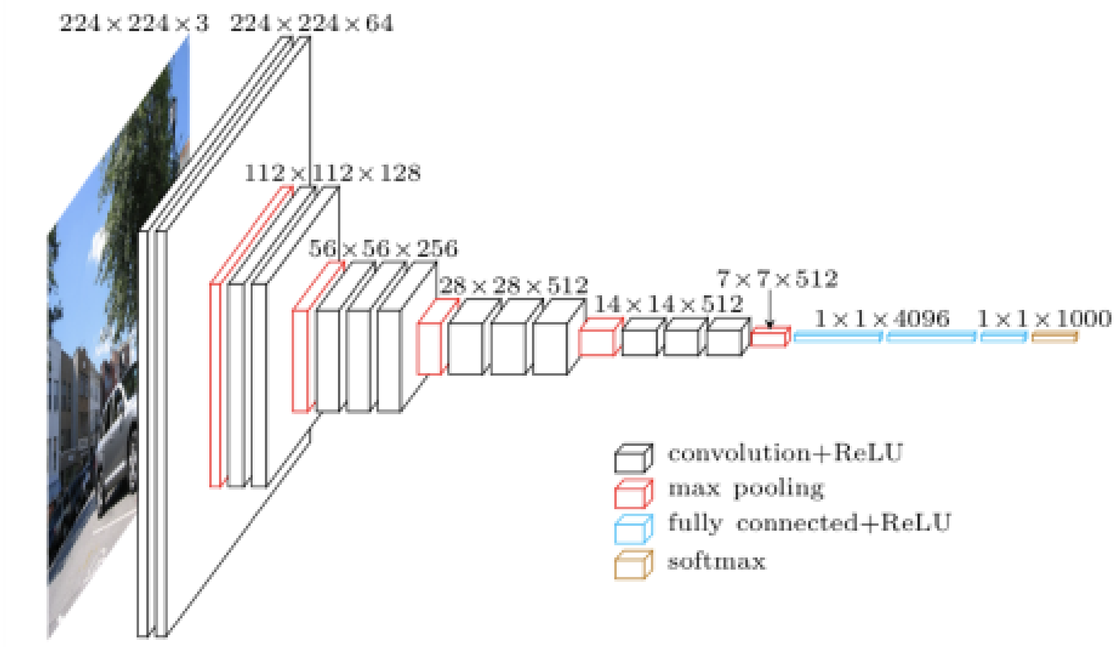


Figure 2: VGG Neural Network Architecture. [4]

- Trained on a dataset consisting of 100 million YouTube videos with 3K labels
- Transfer learning is utilized with pre-trained VGGish weights for initial training and fine-tuning on a different dataset
- Simple classification model achieved an accuracy of 97%

The RELAX method

The RELAX method calculates saliency based on occlusion.

- The cosine similarity measures the similarity between the unmasked, $\mathbf{h}$, and the masked representation, $\bar{\mathbf{h}}$,

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

- The explanations are computed by sampling N masks and computing the sample mean

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

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

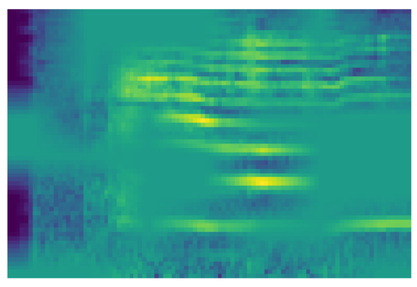
- The uncertainty of the RELAX-score for pixel (i, j)

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

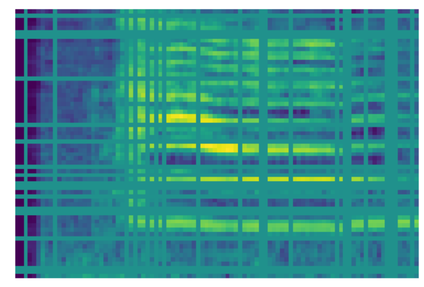
Masking Strategies

The primary focus lies in the masking strategies applied to the spectrograms. Since the project is focusing on spectrogram multiple different masking strategies have been used, which "imposes rules" in time and frequency domain.

- Random masking as in the RELAX paper
- Time-based masking
- Frequency-based masking
- Time- and frequency-based masking

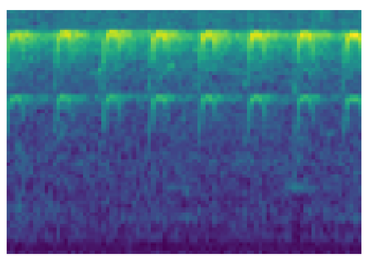


Random masking (From RELAX)

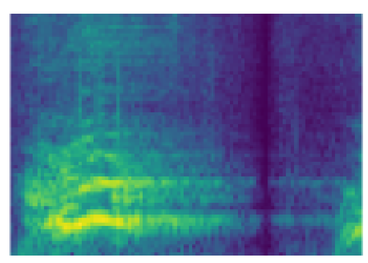


Time- and frequency-based masking

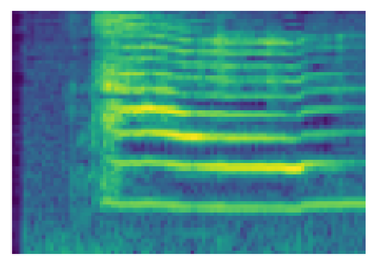
Results



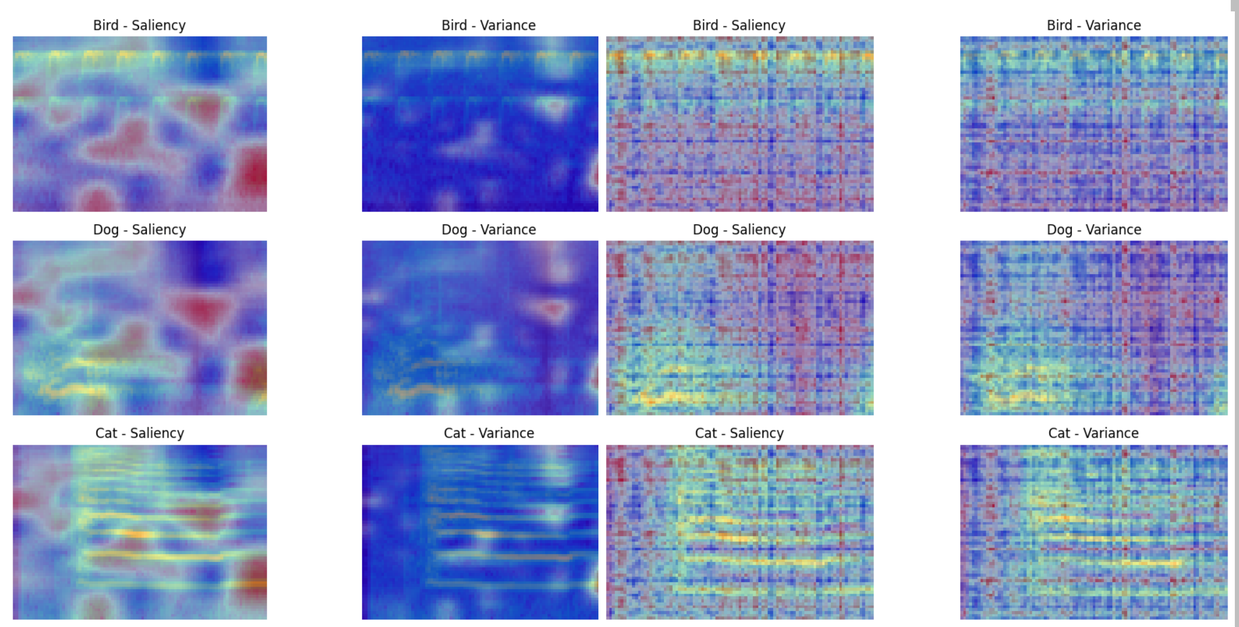
Bird spectrogram



Dog spectrogram



Cat spectrogram

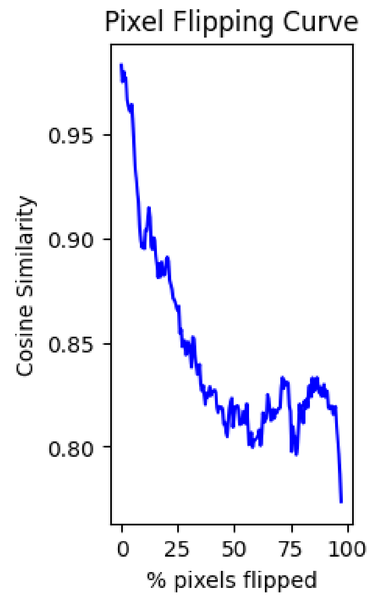


(a) Random masking (From RELAX)

(b) Time- and frequency-based masking

Discussion

- Result verification
  - ▷ Different masking strategies
  - ▷ Pattern formation, localization, and faithfulness
  - ▷ What is important to us?
- Reproducibility issues from images to spectrograms
  - ▷ No definitive way to "mask" out spectrograms
  - ▷ Interpretability issue
  - ▷ Human bias on what is important
  - ▷ Logistic classification on latent space for verification
- Faithfulness - verification of our findings
  - ▷ Pixel flipping - works as "intended"
  - ▷ Registration of spikes



References

[1] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. W. Wilson. CNN architectures for large-scale audio classification. *CoRR*, abs/1609.09430, 2016. URL `http://arxiv.org/abs/1609.09430`.

[2] U. M. I. Lab. Quantus: A framework for understanding and interpreting machine learning models. `https://github.com/understandable-machine-intelligence-lab/Quantus`, Year of Access. Accessed on: Date of Access.

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

[4] VGG. Vgg very deep convolutional networks, 2023. URL `https://viso.ai/deep-learning/vgg-very-deep-convolutional-networks/`.

[5] K. K. Wickstrøm, D. J. Trosten, S. Løkse, A. Boubekki, K. Ø. Mikalsen, M. C. Kampffmeyer, and R. Jenssen. Relax: Representation learning explainability. *International Journal of Computer Vision*, (131):1584–1610, 2023.