Learning methodologies for music and audio data

Emmanouil Benetos
http://www.eecs.qmul.ac.uk/~emmanouilb/
http://machine-listening.eecs.qmul.ac.uk/

Dynamics, Data and Deep Learning Workshop, March 2024

Queen Mary
University of London

Context

centre for digital music

c4dm.eecs.qmul.ac.uk

Context

centre for digital music

c4dm.eecs.qmul.ac.uk

www.aim.qmul.ac.uk

Context

centre for digital music

c4dm.eecs.qmul.ac.uk

www.aim.qmul.ac.uk

CIS centre for intelligent sensing

cis.eecs.qmul.ac.uk

Context

centre for digital music

c4dm.eecs.qmul.ac.uk

www.aim.qmul.ac.uk

CIS centre for intelligent sensing

cis.eecs.qmul.ac.uk

The Alan Turing Institute

www.turing.ac.uk

Talk outline

1. Machine listening
2. Machine listening with limited data
3. Graph neural networks for audio
4. Interpretable machine listening
5. Future perspectives

Machine listening

Machine listening

Machine listening

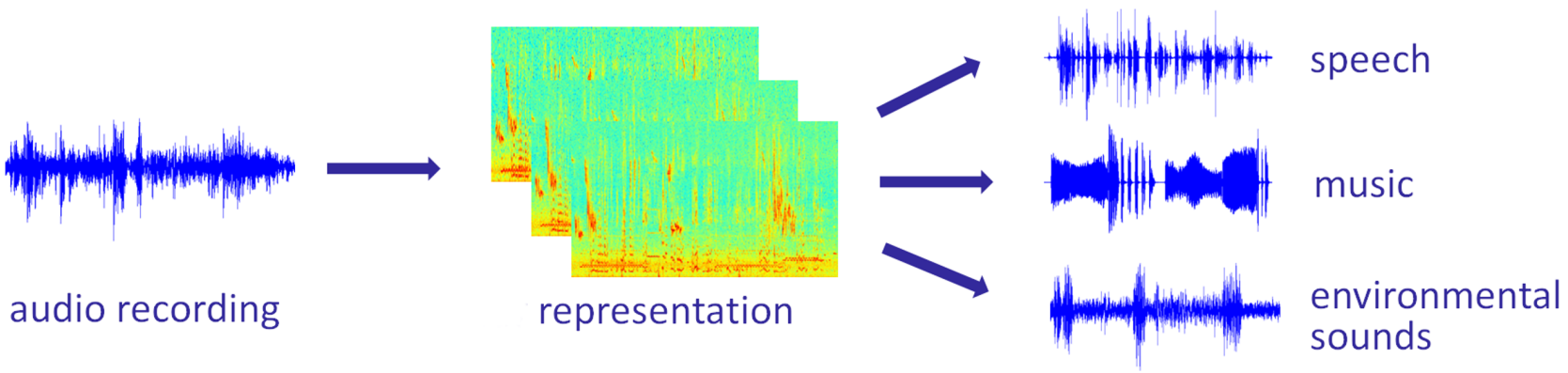
The ability of a machine to interpret and understand audio signals.

Machine listening

Machine listening

The ability of a machine to interpret and understand audio signals.

- **Sounds**: speech, music, environmental/everyday sounds
- **Disciplines**: signal processing, machine learning, acoustics, perception



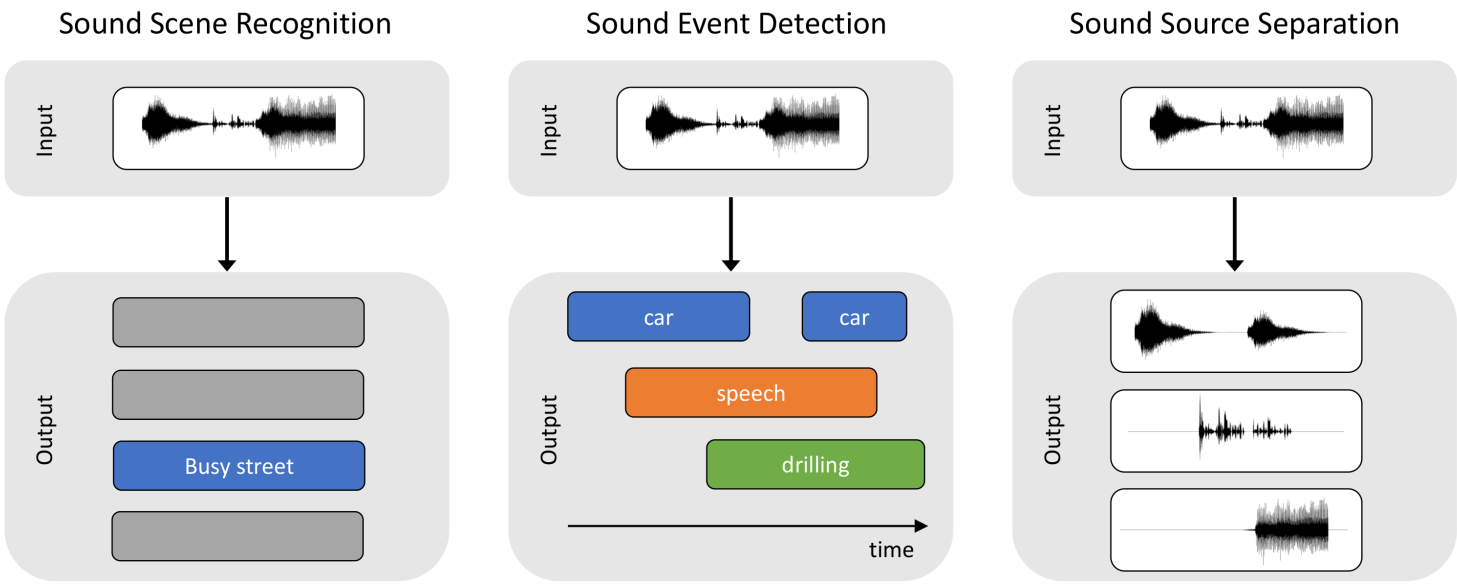
Machine listening for soundscapes

Core problems:

- Sound event detection
- Sound scene recognition
- Source separation
- Noise monitoring/reduction

Applications:

- Smart assistants
- Security
- Industrial monitoring
- Acoustic ecology

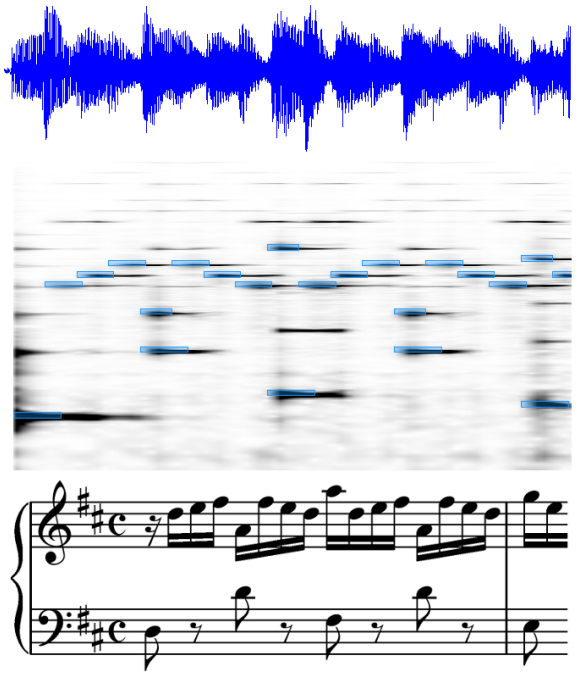


Machine listening for music

Related to the field of **Music Information Retrieval (MIR)**

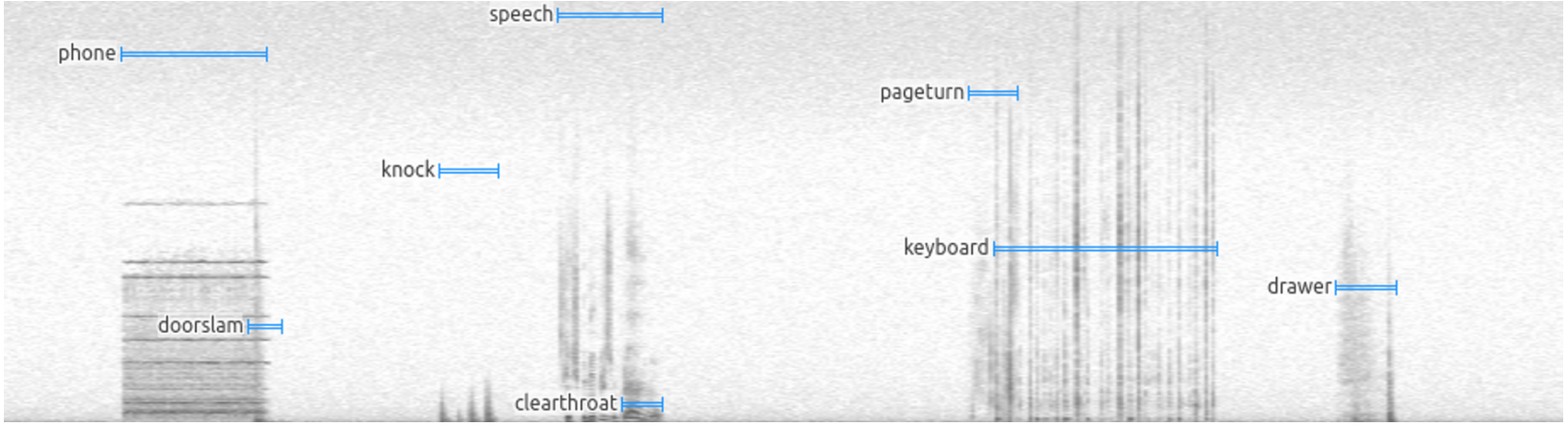
Core problems:

- Music tagging
- Music source separation
- Music transcription
- Audio identification
- & new multimodal music tasks



Challenges in machine listening

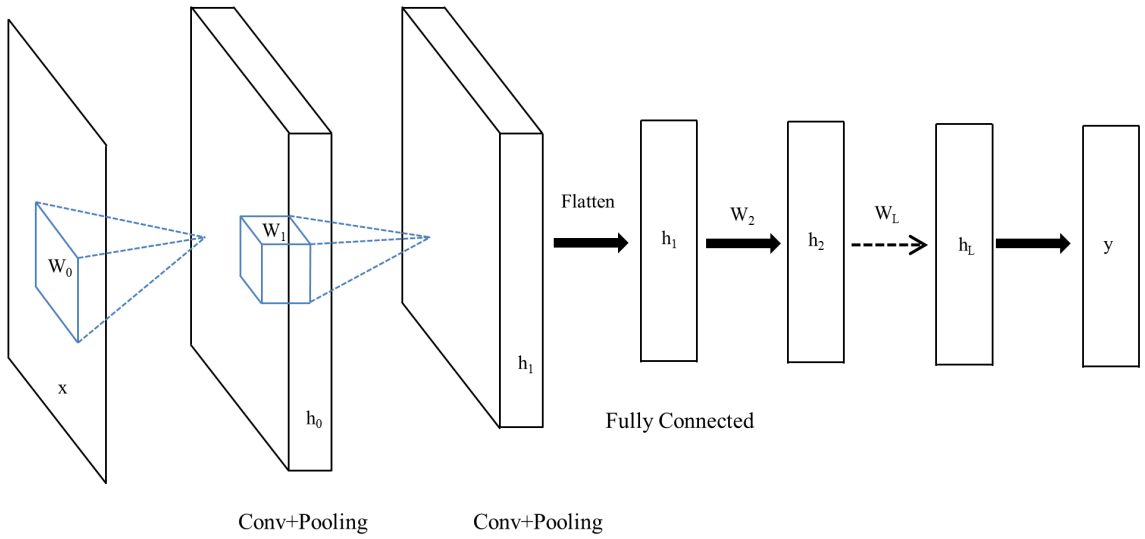
- Multiple overlapping sources
- Data scarcity
- Temporal dependencies
- Noise, distortions and effects
- Unseen domains



Supervised learning for machine listening

Benchmark approaches for machine listening tasks:

- Adopt a **supervised** deep learning approach
- Assume a sufficiently large, **strongly labelled** dataset
- Time-frequency representations or raw waveforms as **input**
- **Building blocks**: feedforward, convolutional & recurrent layers
- **Loss functions**: cross-entropy, MSE



Machine listening with limited data

Domain adaptation for sound recognition

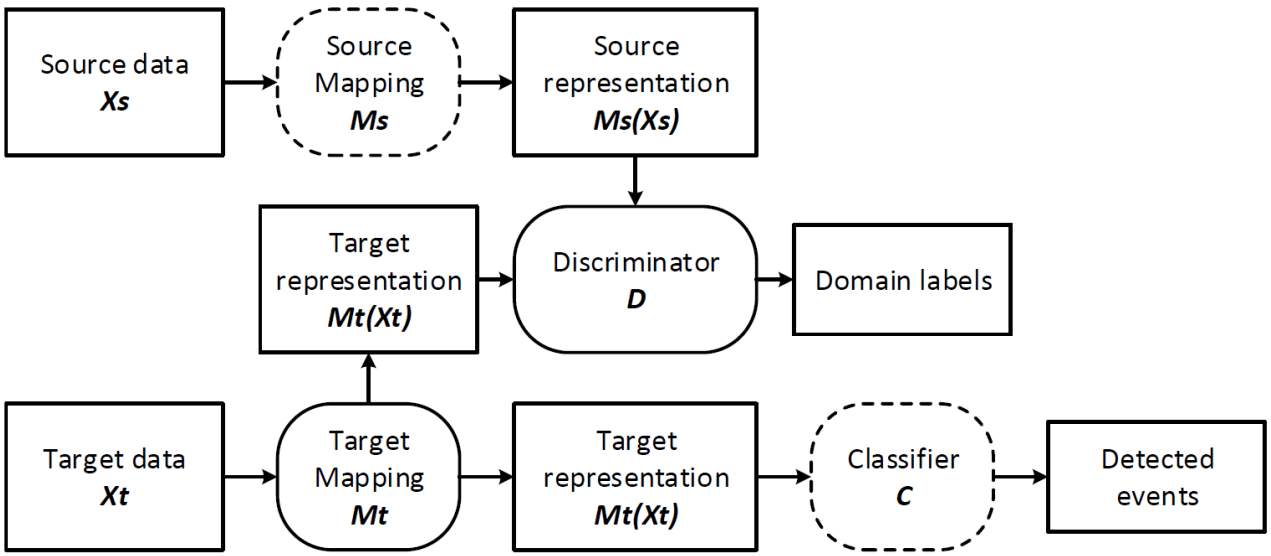
Domain adaptation

Sub-discipline of machine learning which deals with scenarios in which a model trained on a source distribution is used in the context of a different target distribution.

Domain adaptation for sound recognition

Domain adaptation

Sub-discipline of machine learning which deals with scenarios in which a model trained on a source distribution is used in the context of a different target distribution.

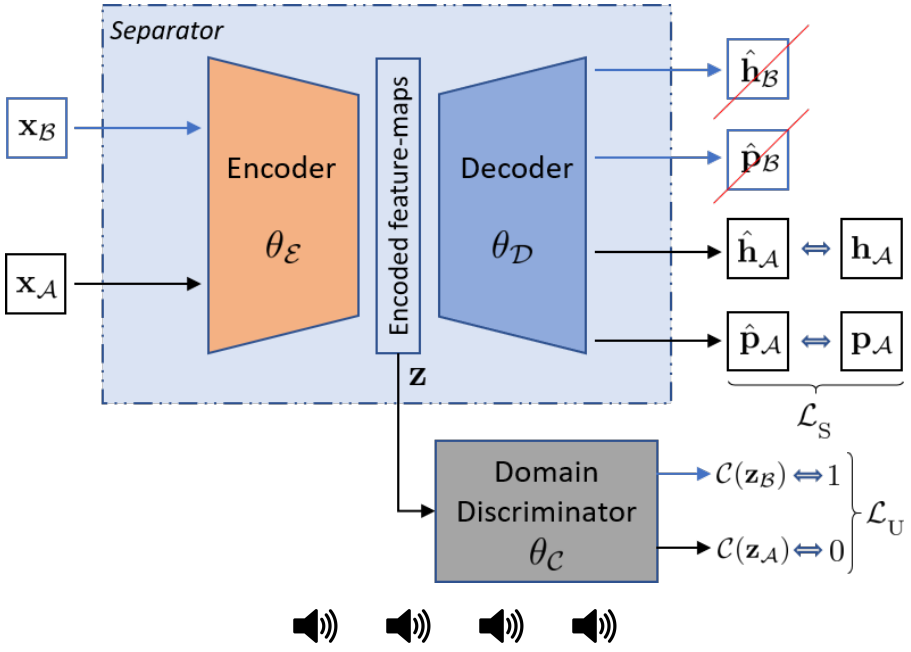


W. Wei, H. Zhu, E. Benetos, and Y. Wang, "A-CRNN: a domain adaptation model for sound event detection", in Proc. IEEE ICASSP, 2020.

Domain adaptation for music source separation

Music source separation system able to adapt to unlabelled mixtures from a new domain.

Framework can be used with any architecture, number of sources, and input representation.



C. Lordelo, E. Benetos, S. Dixon, S. Ahlbäck, and P. Ohlsson, “Adversarial Unsupervised Domain Adaptation for Harmonic-Percussive Source Separation”, IEEE Signal Processing Letters, 28:81-85, 2021.

Few-shot learning for audio classification

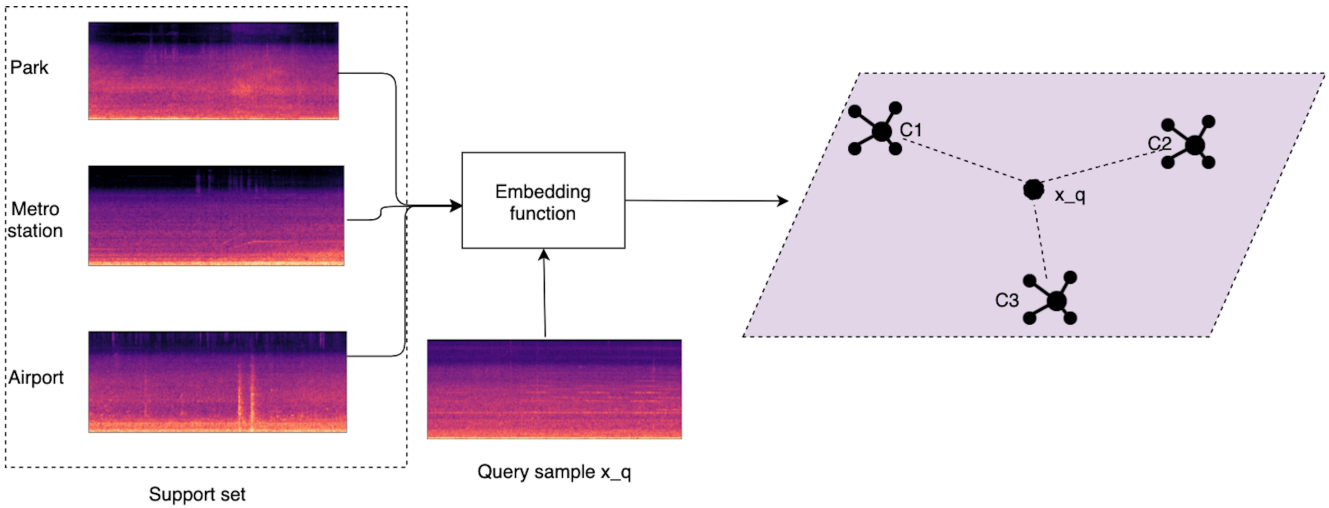
Few-shot learning

Learning from a limited number of labelled examples.

Few-shot learning for audio classification

Few-shot learning

Learning from a limited number of labelled examples.



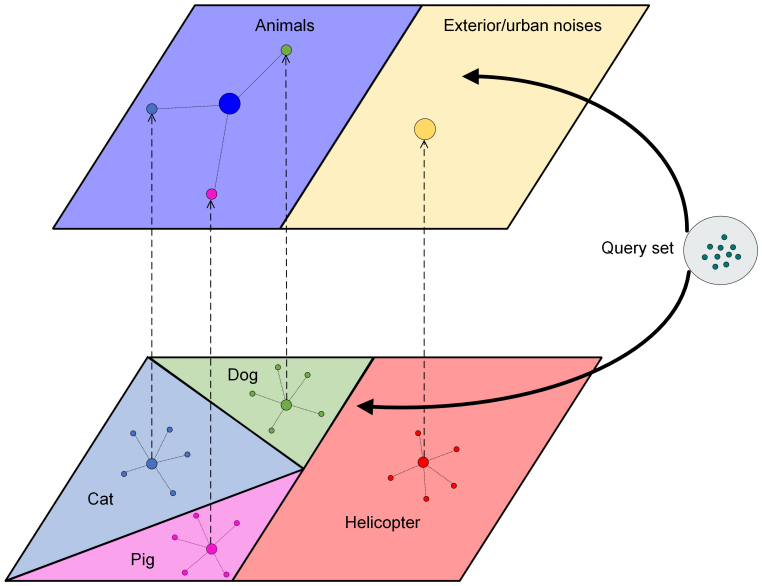
Each class prototype c_k is the mean of the embedded support points x_i belonging to its class: $c_k = \frac{1}{|S_k|} \sum_{(x_i) \in S_k} f_\phi(x_i)$

S. Singh, H. L. Bear, and E. Benetos, "Prototypical networks for domain adaptation in acoustic scene classification", in Proc. ICASSP, 2021.

Few-shot learning for sound recognition

Proposing a **hierarchical prototypical network** to leverage knowledge rooted in audio taxonomies.

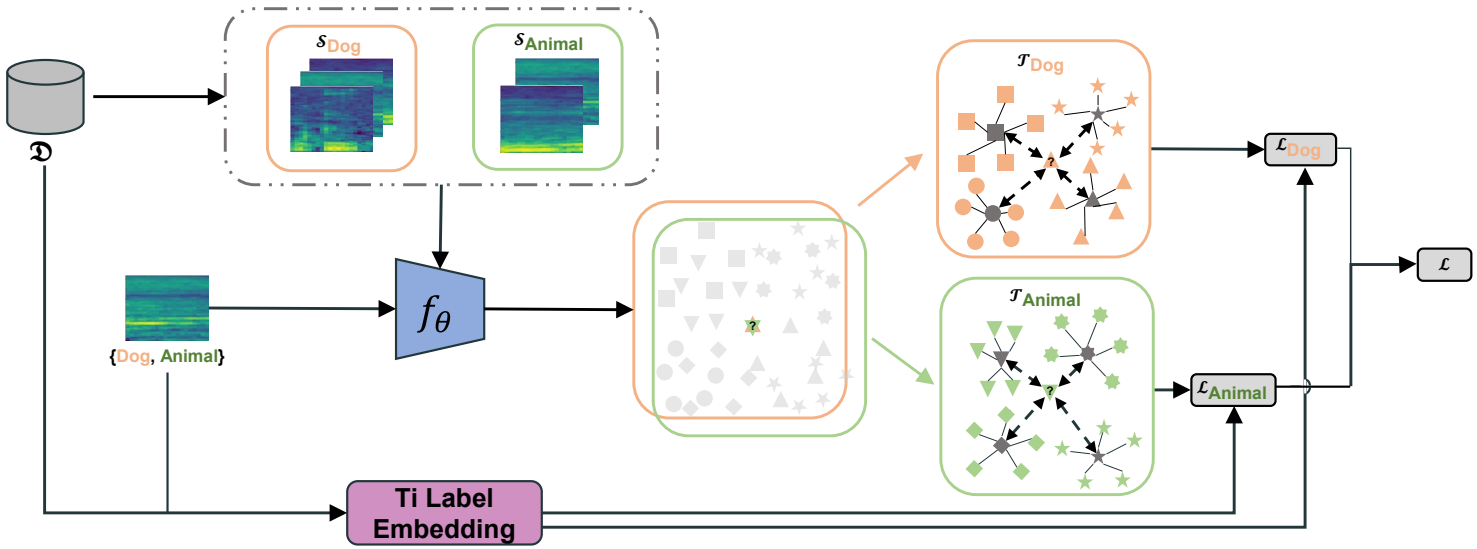
- Prototypes at the lower level: $c_k^{(0)} = \frac{1}{|S_k|} \sum_{(x_i) \in S_k} f_\phi(x_i)$
- Prototypes at a higher level h in the taxonomy: $c_j^{(h)} = \frac{1}{|C_k^{(h)}|} \sum_{c_j^{(h)} \in C_j^{(h)}} c_j^{(h-1)}$



J. Liang, H. Phan, and E. Benetos, "Leveraging label hierarchies for few-shot everyday sound recognition", in Proc. DCASE, 2022.

Few-shot learning for sound recognition

- Extending hierarchical prototypical networks to multi-label classification problems.
- Converts a multi-label classification problem to multiple single-label tasks & incorporates taxonomy knowledge in the training objective.

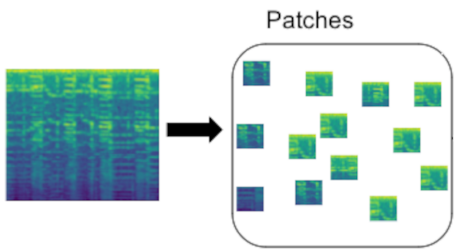


J. Liang, H. Phan, E. Benetos, "Learning from taxonomy: multi-label few-shot classification for everyday sound recognition", in IEEE ICASSP, 2024.

Graph neural networks for audio

Graph neural networks for audio

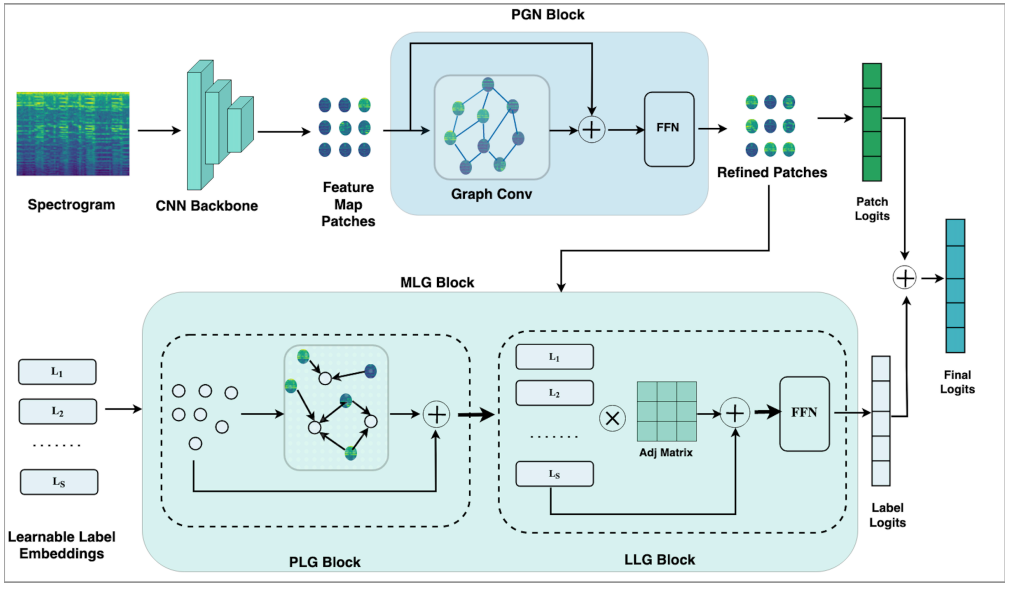
- Despite stacking multiple layers, the receptive field of convolutional layers remains severely limited.
- Attention mechanisms are able to map global context, but are not flexible enough to capture irregular audio objects.
- Treating time-frequency representations as graph structures instead



Graph neural networks for audio

Proposed model converts spectrogram into a graph structure, and maps relationships between class features and corresponding spectrogram regions.

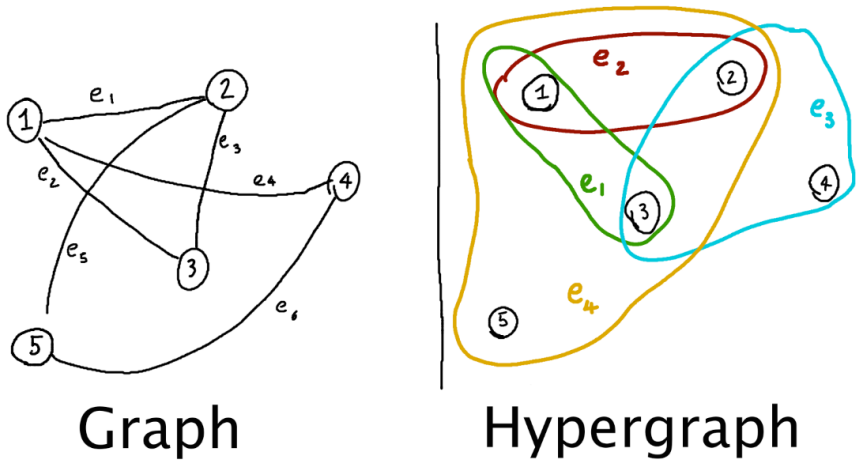
Comparable results with non-graph models, with significantly lower number of learnable parameters.



S. Singh, C. J. Steinmetz, E. Benetos, H. Phan, and D. Stowell, "ATGNN: audio tagging graph neural network", IEEE Signal Processing Letters, 2024.

Graph neural networks for audio

- Model higher level representations beyond pairwise relationships
- Learn audio “object” representation
- **Hypergraphs** - Extension of graphs to higher level representations
- Single hyperedge can cover multiple nodes



Graph Hypergraph

Interpretable machine listening

Interpretable machine listening

Interpretability in machine learning

The ability to explain or present the behaviour of a machine learning model in understandable terms to a human.

Interpretable machine listening

Interpretability in machine learning

The ability to explain or present the behaviour of a machine learning model in understandable terms to a human.

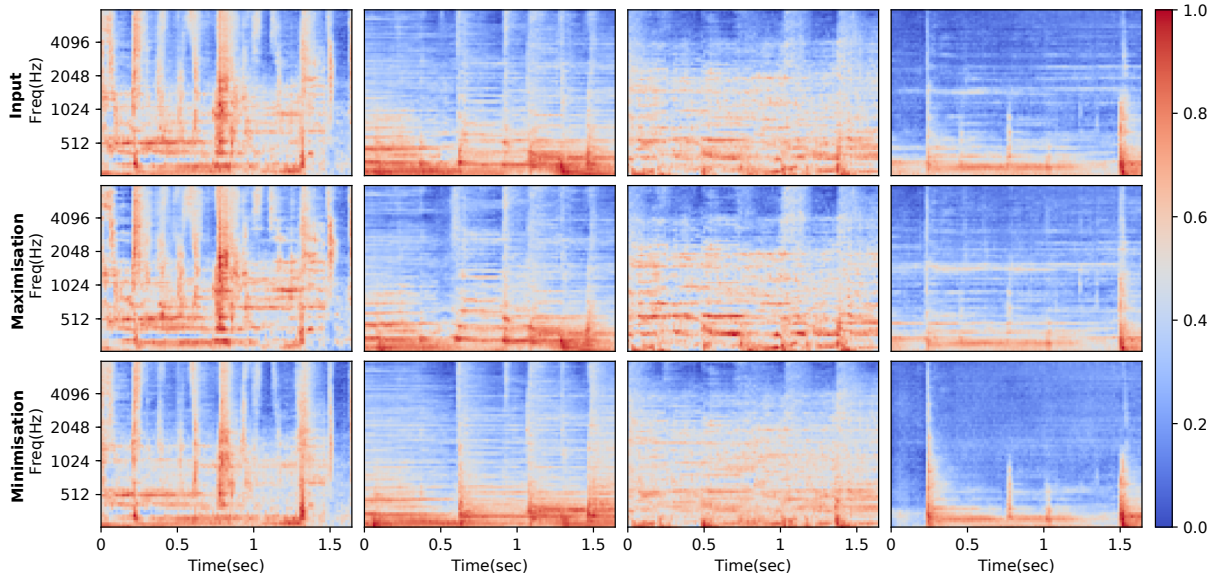


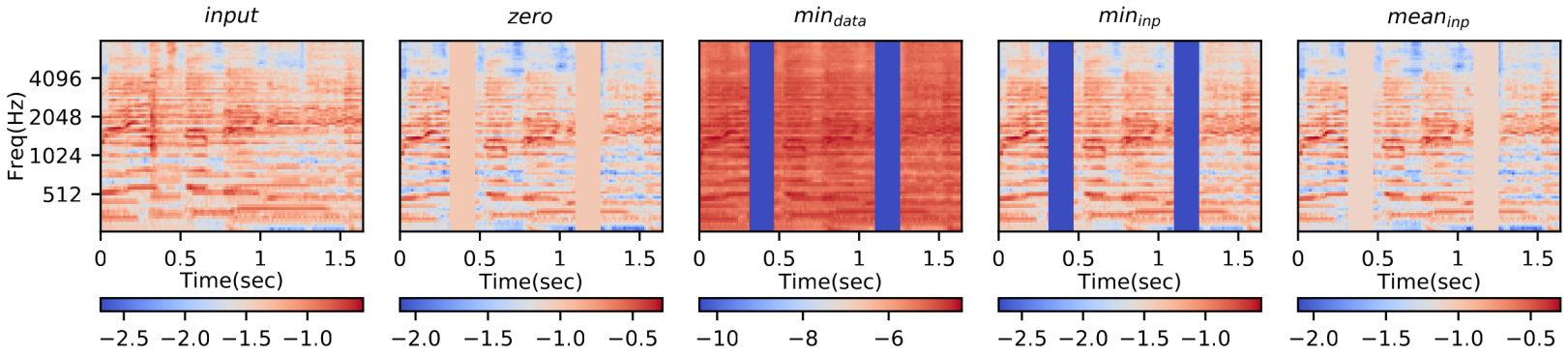
Figure: examples synthesised by maximally and minimally activating the output neuron of a singing voice detection model.

S. Mishra, D. Stoller, E. Benetos, B. Sturm and S. Dixon, “GAN-based generation and automatic selection of explanations for neural networks”, in ICLR 2019 Workshop on Safe Machine Learning (SafeML), 2019.

Interpretable machine listening

Local explanations for machine listening tasks by perturbations on the input representation.

Investigating the reliability of explanations for machine listening tasks

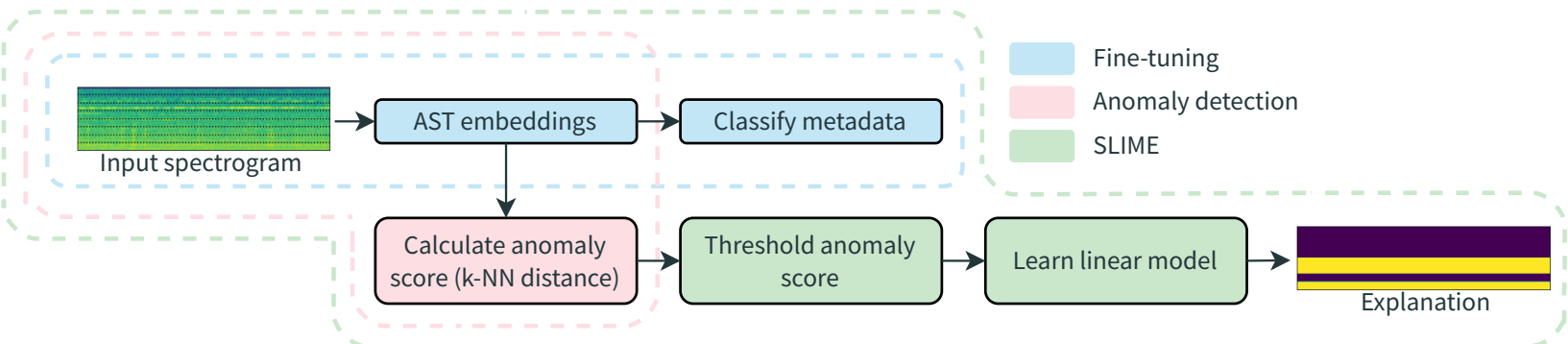


S. Mishra, E. Benetos, B. L. Sturm and S. Dixon, “Reliable local explanations for machine listening”, in IJCNN, 2020.

Explaining anomalous sound detectors

The task of deciding whether a sound produced from an object is normal or anomalous; only inliers are available for training

Local explanations show that deep learning models rely on higher frequencies to distinguish anomalies from inliers, and can be replaced by simple high-pass signal processing based models

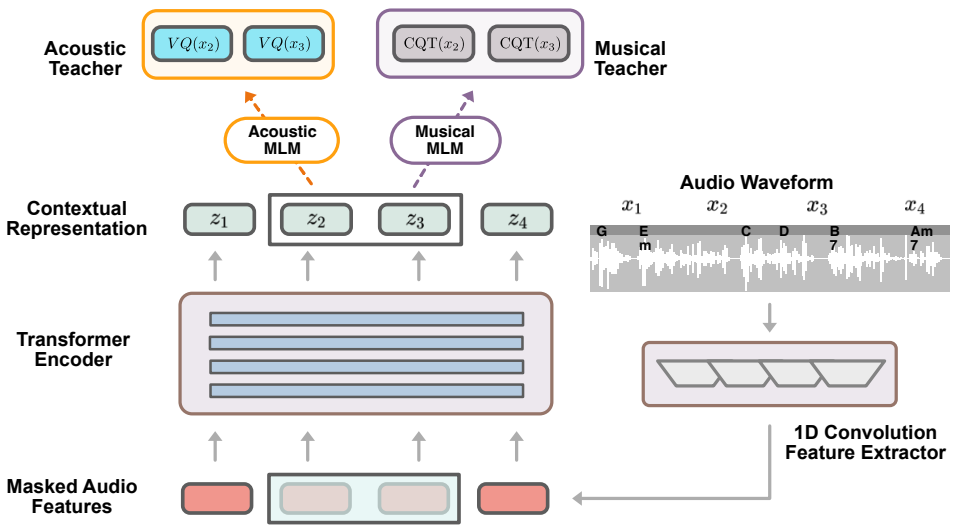


K. T. Mai, T. Davies, L. D. Griffin, and E. Benetos, “Explaining the decisions of anomalous sound detectors”, in DCASE, 2022.

Future perspectives

Future perspectives

- Incremental learning for audio and music
- Multimodal AI for music and audio understanding: audio, scores, video, images, tags, captions...
- Self-supervised learning
- Audio & music foundation models



Many thanks to

- Simon Dixon
- Jinhua Liang
- Carlos Lordelo
- Yinghao Ma
- Kimberly Ton Mai
- Saumitra Mishra
- Huy Phan
- Shubhr Singh
- Dan Stowell
- Bob L. Sturm
- Wei Wei

SUPPORT:

UKRI Engineering and Physical Sciences Research Council

Royal Academy of Engineering