Explainable Machine Learning and its applications to Machine Listening

SAUMITRA MISHRA$^1$, BOB L. STURM$^2$, EMMANOUIL BENETOS$^1$, AND SIMON DIXON$^1$

$^1$Centre for Digital Music, School of Electronic Engineering and Computer Science, Queen Mary University of London
$^2$School of Electronic Engineering and Computer Science, Royal Institute of Technology KTH, Sweden
Motivation
Machine Learning (specifically deep learning) + Big data + High computational power are state-of-the-art in many applications. **But,**

- The black-box nature of algorithms,
- susceptibility to adversarial attacks,
- lack of mathematical and empirical understanding

has raised concerns.

---

https://www.wired.com/story/ai-has-a-hallucination-problem-thats-proving-tough-to-fix/

https://www.telegraph.co.uk/technology/2017/08/01/facebook-shuts-robots-invent-language/


Is a machine learning model **Trustworthy?** **Robust?** **Fair?** ...?

How can we address the above concerns?

Is a machine learning model  **Trustworthy?**  **Robust?**  **Fair?**  ...?

How can we address the above concerns?

By bringing **Interpretability** to machine learning models
   Or
   Explainable AI
       Or
   Interpretable Machine Learning

---

Is a machine learning model Trustworthy? Robust? Fair? ...?

How can we address the above concerns?

By bringing Interpretability to machine learning models
Or
Explainable AI
Or
Interpretable Machine Learning

Methods to understand model behaviour
Train Inherently Interpretable Models
- Decision trees
- Sparse linear models
- Rule-based models

Interpretable Machine Learning
**Interpretable Machine Learning**

- **Train Inherently Interpretable Models**
  - Decision trees
  - Sparse linear models
  - Rule-based models

- **Limitations**
  - Uninterpretable Features
  - Poor performance on high-dimensional data.
  - Difficult to optimize.
Interpretable Machine Learning

- Train Inherently Interpretable Models
  - Decision trees
  - Sparse linear models
  - Rule-based models

Limitations
- Uninterpretable Features
- Poor performance on high-dimensional data.
- Difficult to optimize.

Explain a model (global analysis)
Explain a prediction (Local Analysis)
Post-hoc Analysis of Pre-trained models

Transparent Machine Learning

Train Inherently Interpretable Models
- Decision trees
- Sparse linear models
- Rule-based models

Limitations
- Uninterpretable Features
- Poor performance on high-dimensional data.
- Difficult to optimize.

Explain a model (global analysis)
- Feature inversion
- Activation maximisation
  - Synthetic
  - Dataset-based

Post-hoc Analysis of Pre-trained models

Explain a prediction (Local Analysis)
Transparent Machine Learning

Train Inherently Interpretable Models
- Decision trees
- Sparse linear models
- Rule-based models

Limitations
- Uninterpretable Features
- Poor performance on high-dimensional data.
- Difficult to optimize.

Explain a model (global analysis)
- Feature inversion
- Activation maximisation
  - Synthetic
  - Dataset-based

Explain a prediction (Local Analysis)
- Sensitivity Analysis
  - Occlusion
- Gradient-based saliency maps
- Function Decomposition
  - Layer-wise relevance propagation
  - Deconvolution network
- Miscellaneous
  - Combine global approximation with local sensitivity analysis - LIME

Post-hoc Analysis of Pre-trained models
Feature Inversion

Feature inversion (or inverting a feature vector) involves mapping (in some way) a feature vector from a layer, back to the input space (e.g., image, time-frequency spectrogram).
Feature Inversion

Feature inversion (or inverting a feature vector) involves mapping (in some way) a feature vector from a layer, back to the input space (e.g., image, time-frequency spectrogram).

How can we use this idea of feature inversion to understand a model?
Key Idea

Discriminative training forces each hidden layer of a deep discriminative model to only preserve information relevant to the discrimination task.
Key Idea

Discriminative training forces each hidden layer of a deep discriminative model to only preserve information relevant to the discrimination task.

Inversion of features (e.g., features maps) generated at a layer (e.g., convolutional layer) back to the input space (e.g., pixel space) will assist in understanding information a model preserves at that layer.
Deep Vocal Detector

Mel Spectrogram

Conv1 → Conv2 → MP3 → Conv4 → MP6 → FC7 → FC8 → P(vocal) = 0.1

Feature Inversion

Inverted Representations

Results

- FC8 does not preserve any temporal and harmonic information, but the reconstructions from shallower layers are visually similar to the input.
- Inverted representations from FC8 suggest that the SVD model learns a class-decision function in this layer.
- Deeper layers capture more invariances from data than shallow layers.
- The above results generalize across datasets.
Take Away Points

- Relying just on performance metrics for model selection may lead to selection of suboptimal models.
- Combining performance metrics with interpretable explanations may provide more insight into model behaviour, leading to the development and selection of trustworthy models.
- There exist several ways to analyse the behaviour of machine learning models
  - Using inherently Interpretable models
  - Using post-hoc methods to analyse a pre-trained model.
THANK YOU