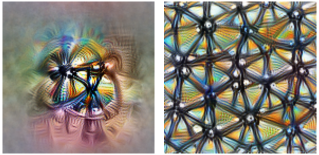
Interpretable Machine Learning

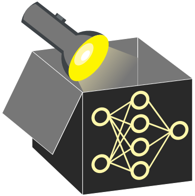
Post-hoc Methods for Neural Networks



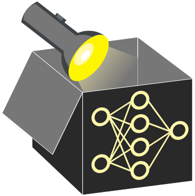
Learning goals

- Interpretability in neural networks
- Landscape of interpretability
- The difference between feature visualization and feature attributions

NEURAL NETWORKS AS COMPLEX ML MODELS

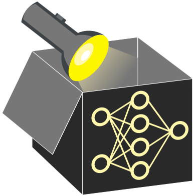


NEURAL NETWORKS AS COMPLEX ML MODELS

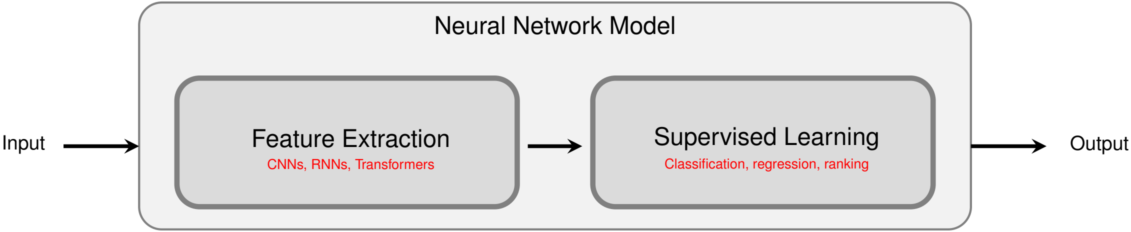


- Neural networks are over parameterised
  - Vision models and Language models routinely have > millions of params
  - Sometimes #parameters > #input instances
  - Which and how do the features, parameters, training instances contribute towards the final decision ?

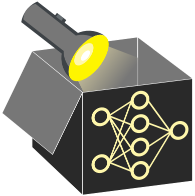
NEURAL NETWORKS AS COMPLEX ML MODELS



- Neural networks are over parameterised
  - Vision models and Language models routinely have > millions of params
  - Sometimes #parameters > #input instances
  - Which and how do the features, parameters, training instances contribute towards the final decision ?

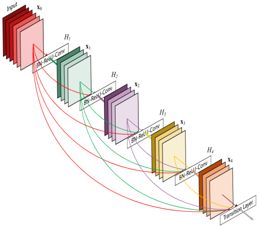


NEURAL NETWORKS AS COMPLEX ML MODELS

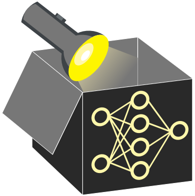


- Neural networks are compositional and non-linear systems
  - The success of neural networks is due to their depth
    - Depth results in compositional behaviour
  - Non-linearity between layers helps capture non-linear relationships

- Depth and non-linearity leads to lack of interpretability

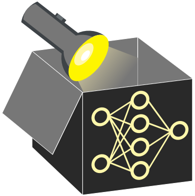


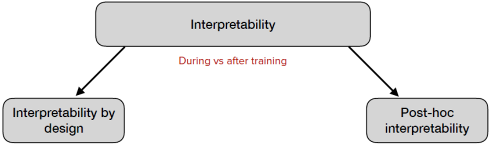
MODEL-SPECIFIC INTERPRETABILITY



- What types of neural models are out there ?
  - For vision: Convolutional Neural Nets
  - For language, speech: Recurrent Neural Nets, Transformer Models
  - For recommendation systems, ranking: Factorization-based Models, Embeddings models
- Each of the domains have their challenges and have developed specific approaches for interpretability
  - We will focus on first principles that can be applied to most models
  - We will discuss adaptations to each data modality as and when required

INTERPRETABILITY LANDSCAPE IN NEURAL NETWORKS





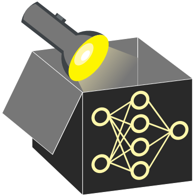
Interpretability

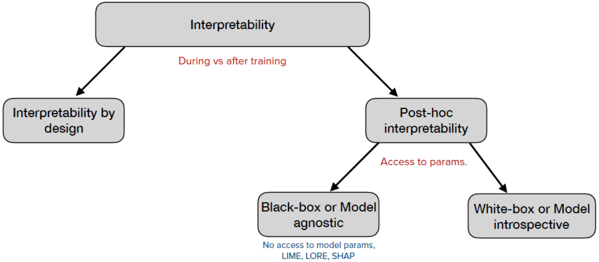
During vs after training

Interpretability by design

Post-hoc interpretability

INTERPRETABILITY LANDSCAPE IN NEURAL NETWORKS





Interpretability

During vs after training

Interpretability by design

Post-hoc interpretability

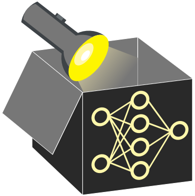
Access to params.

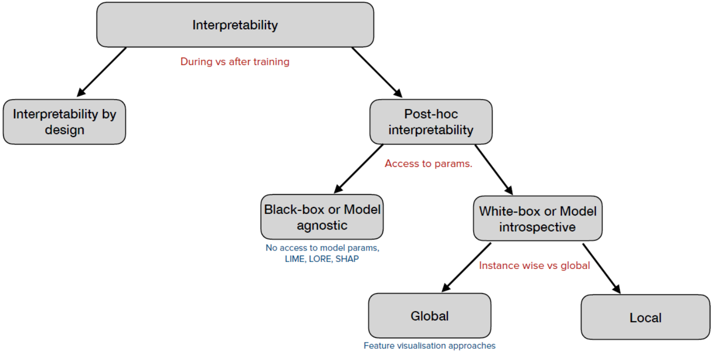
Black-box or Model agnostic

No access to model params, LIME, LORE, SHAP

White-box or Model introspective

INTERPRETABILITY LANDSCAPE IN NEURAL NETWORKS





Interpretability

During vs after training

Interpretability by design

Post-hoc interpretability

Access to params.

Black-box or Model agnostic

No access to model params, LIME, LORE, SHAP

White-box or Model introspective

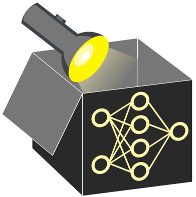
Instance wise vs global

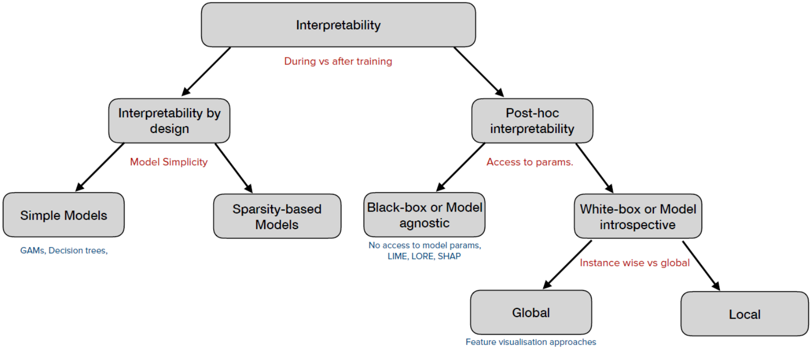
Global

Feature visualisation approaches

Local

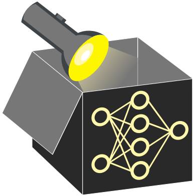
INTERPRETABILITY LANDSCAPE IN NEURAL NETWORKS

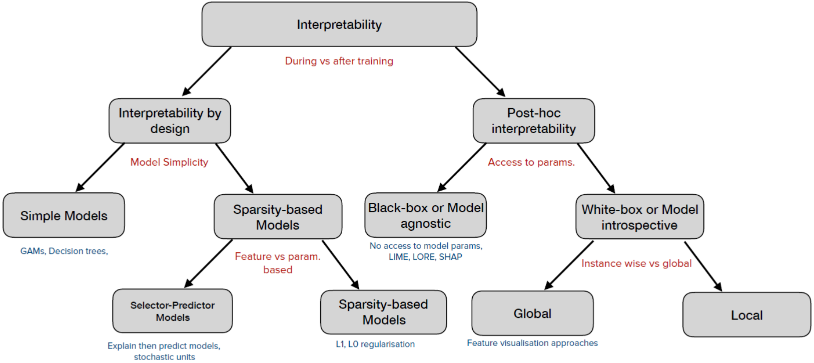




Interpretability
During vs after training
Interpretability by design
Post-hoc interpretability
Model Simplicity
Access to params.
Simple Models
Sparsity-based Models
Black-box or Model agnostic
White-box or Model introspective
GAMs, Decision trees,
No access to model params, LIME, LORE, SHAP
Instance wise vs global
Global
Local
Feature visualisation approaches

INTERPRETABILITY LANDSCAPE IN NEURAL NETWORKS





Interpretability

During vs after training

Interpretability by design

Post-hoc interpretability

Model Simplicity

Access to params.

Simple Models

Sparsity-based Models

Black-box or Model agnostic

White-box or Model introspective

GAMs, Decision trees,

Feature vs param. based

No access to model params, LIME, LORE, SHAP

Instance wise vs global

Selector-Predictor Models

Sparsity-based Models

Global

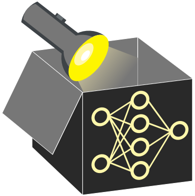
Local

Explain then predict models, stochastic units

L1, L0 regularisation

Feature visualisation approaches

HOW CAN WE INTERPRET NEURAL MODELS ?



- Feature visualization: Visualizing components of the neural networks
  - Activations of neurons
  - Attention values
  - Gradient flow
- Feature attributions: relevant input features
  - Which input features are responsible for the given decision ?
  - Sensitivity analysis using gradient-based methods
  - Using black-box methods like LIME, SHAP, etc.

HOW CAN WE INTERPRET NEURAL MODELS ?

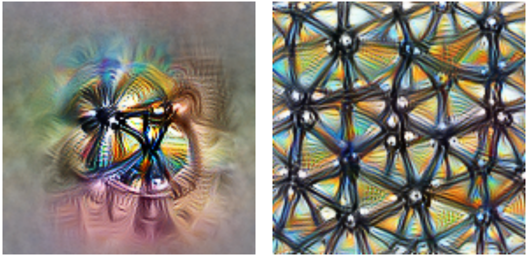


Figure: Feature visualization: Visualizing components of the neural networks

HOW CAN WE INTERPRET NEURAL MODELS ?

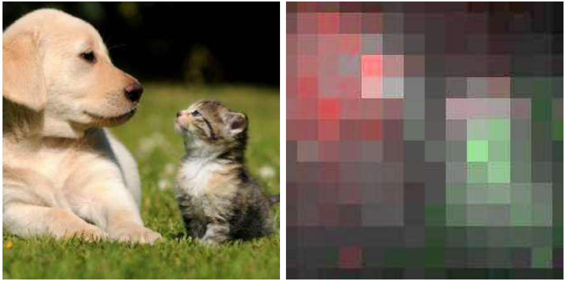


Figure: Feature attributions: relevant input features