Interpretable ML

An Introduction

Fahimeh Hosseini
Ali Almasi

Motivation and Definition

Motivation

What vs. Why

"For starters, I think we should find out who made the coffee that day!"

Motivation What vs. Why

A machine learning model performs well.

*Can **we just trust the model** and ignore **why** it made a certain decision?*

Motivation What vs. Why

The need for interpretability arises from an incompleteness in problem formalization. (Doshi-Velez and Kim 2017)

Motivation What vs. Why

- Human curiosity and learning
- Finding meaning in the world
- Some tasks require safety measures
- Detecting bias
- Social Acceptance of machines

Interpretation Definition

General notion of interpretation:

To extract information (of some form) from data.

In ML:

*[The] extraction of **relevant** knowledge from a machine-learning model concerning relationships either contained in data or learned by the model.*

It is also known as:

explainable ML, intelligible ML, or transparent ML.

Interpretation Definition

Other definitions:

- *Interpretability is the degree to which a human can understand the cause of a decision.*
- *Interpretability is the degree to which a human can consistently predict the model's result.*

Background

- Providing an overview of different interpretation methods
- Evaluating interpretations and what properties should be satisfied

The previous works do not address interpretable machine learning as a whole!

Other Related Areas

- Considering bias and fairness in ML models
- Psychology
- Causal Inference
- Stability

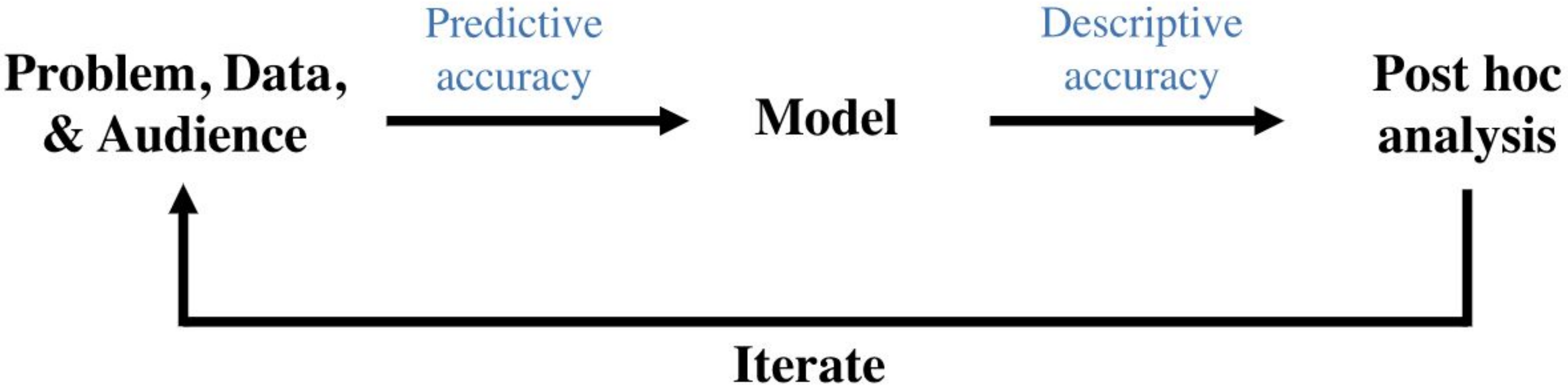
Taxonomy of Interpretation Methods

Classifying Interpretation Methods

Methods for machine learning interpretability can be classified according to various criteria:

- Model-based or Post hoc
- Result of the interpretation method
- Model-specific or model-agnostic
- Local or global

Data-Science Life Cycle



Model-based (Intrinsic) Interpretability

Definition:

The interpretability used in the modeling stage is called model-based interpretability.

Focuses on the construction of models that readily provide insight into the relationships they have learned.

Post Hoc Interpretability

Definition:

The interpretation we do in the post hoc analysis stage is called post hoc interpretability.

Takes a trained model as input and extract information about what relationships the model has learned.

Which one of Interpretation Methods?!

The PDR Desiderata for Interpretations

Accuracy:

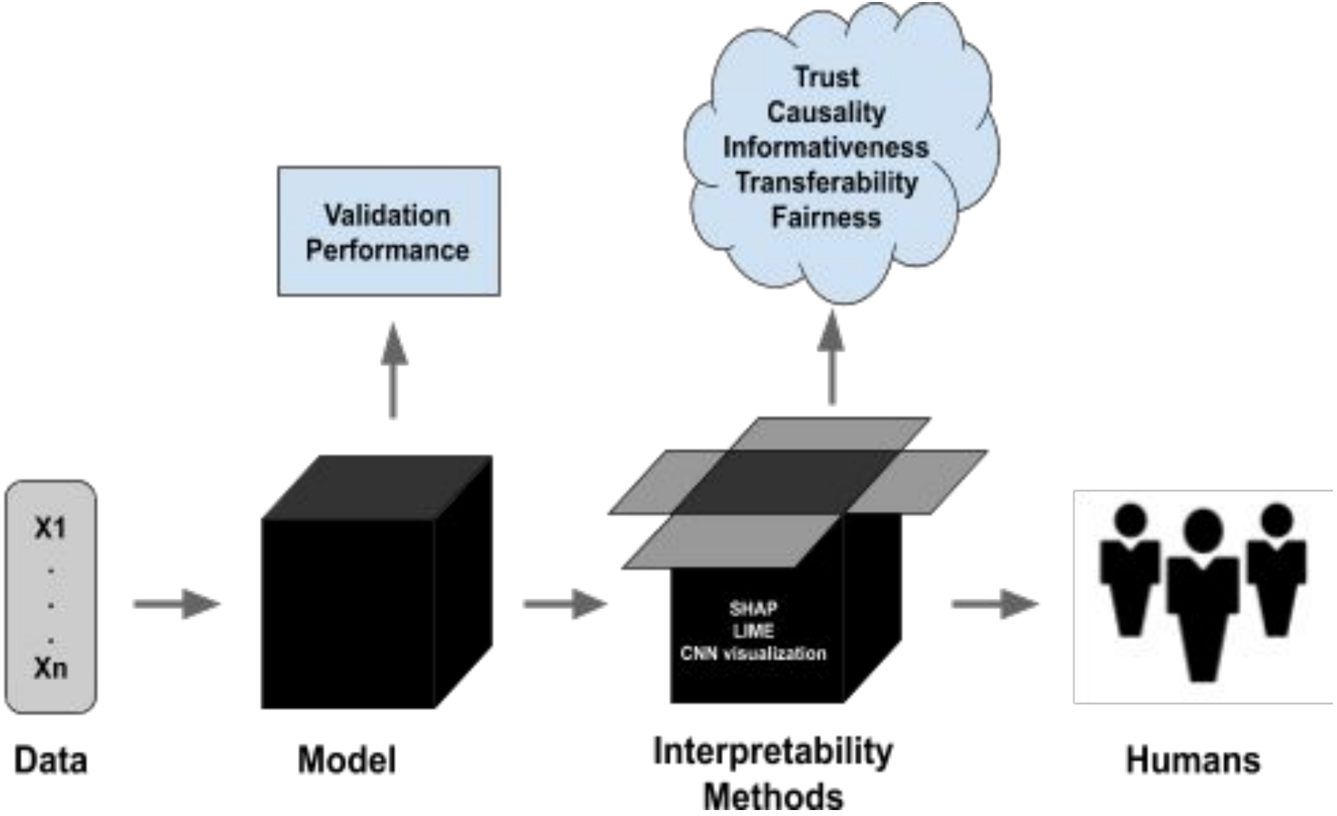
- **Predictive accuracy**
  - The data used to check for predictive accuracy must resemble the population of interest
  - The distribution of predictions matters.
  - Stability matters.
- **Descriptive accuracy**
  The degree to which an interpretation method objectively captures the relationships learned by machine-learning models.

Relevancy:

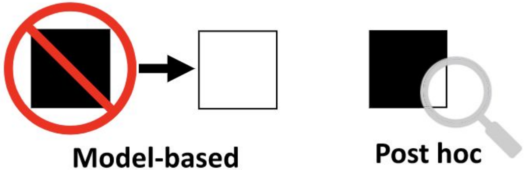
We define an interpretation to be relevant if it provides insight for a particular audience into a chosen domain problem.

It often plays a key role in determining the trade-off between predictive and descriptive accuracy.

The PDR Desiderata for Interpretations



The Impact of Interpretability Methods on Accuracies



	Model-based interpretability	**Post hoc interpretability**
Predictive Accuracy	Generally unchanged or decrease (data-dependent)	No Effect
Descriptive Accuracy	Increase	Increase

Model-Based Interpretability

Model-Based Interpretability, an Overview

- Construction of models that readily provide insight into the relationships they have learned.

- **Desiderata according to PDR framework (ordered by priority):**
  - Predictive accuracy
  - Descriptive accuracy
  - Relevancy

CHALLENGE!

Come up with models that are simple enough to be easily understood, while maintaining high predictive accuracy.

Model-Based Interpretability

- Sparsity
- Simultability
- Modularity
- Feature Engineering

Sparsity

- Often useful in high-dimensional problems
- If we know that the underlying relationship is based on a sparse set of signals…
  Why not ***"impose"*** sparsity?
- Can increase both predictive and descriptive accuracy
- However… Check out for stability of parameters!

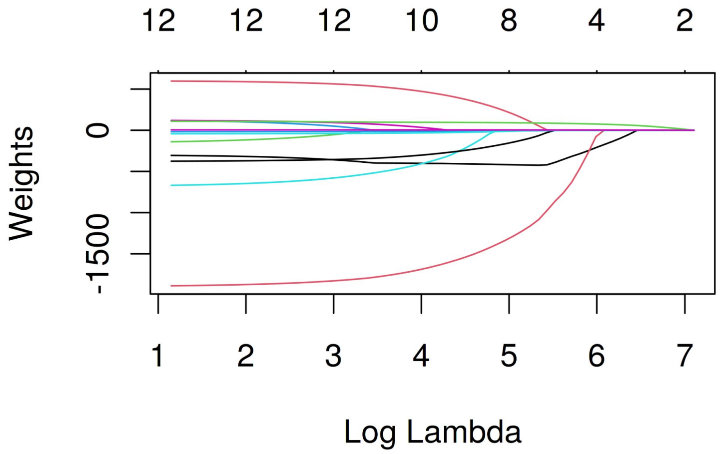
Example of Sparsity: LASSO

- Classic Linear Regression:

$$min_{\beta}\left(\frac{1}{n}\sum_{i=1}^{n}(y^{(i)}-x_i^T\beta)^2\right)$$

- Lasso:Impose sparsity with a regularization term

$$min_{\beta}\left(\frac{1}{n}\sum_{i=1}^{n}(y^{(i)}-x_i^T\beta)^2+\lambda||\beta||_1\right)$$

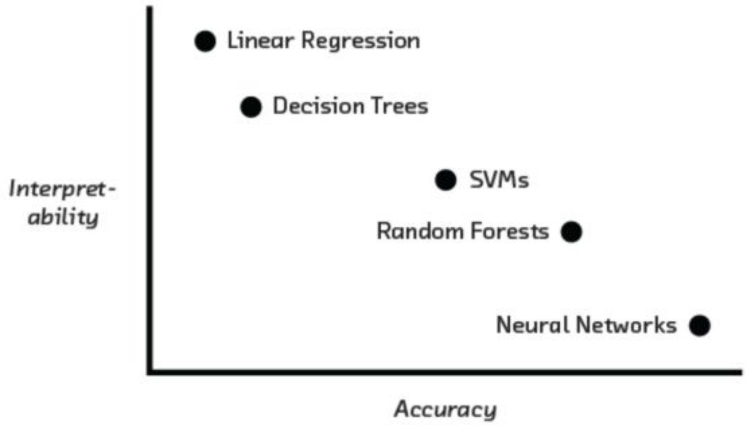


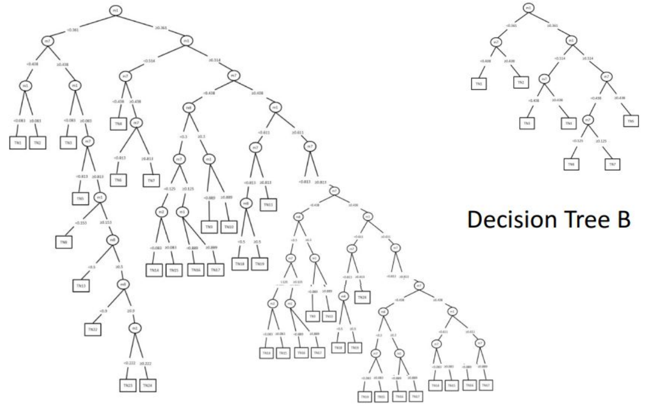
Simulatability

- Use models that human is able to internally simulate and reason about its entire decision-making process (i.e., how a trained model produces an output for an arbitrary input).

- Useful when the number of features is low and the underlying relationship is simple.
  - Decision Trees
  - Rule-Based Learning
  - Linear Regression

- As the complexity of the model increases, it becomes increasingly difficult for a human to internally simulate.





Decision Tree B

Decision Tree A

Modularity

An ML model is modular if a meaningful portion(s) of its prediction-making process can be interpreted independently.

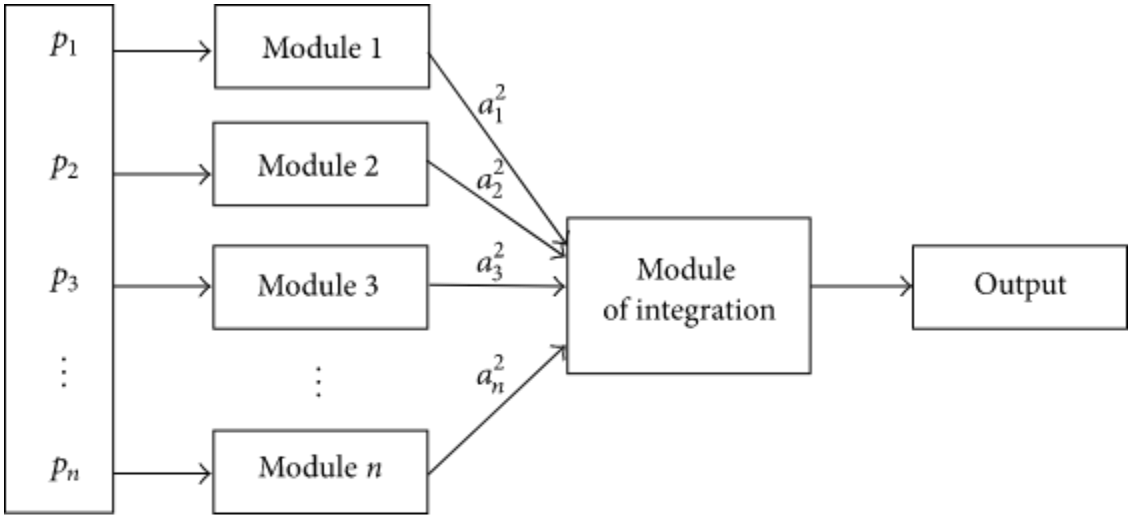
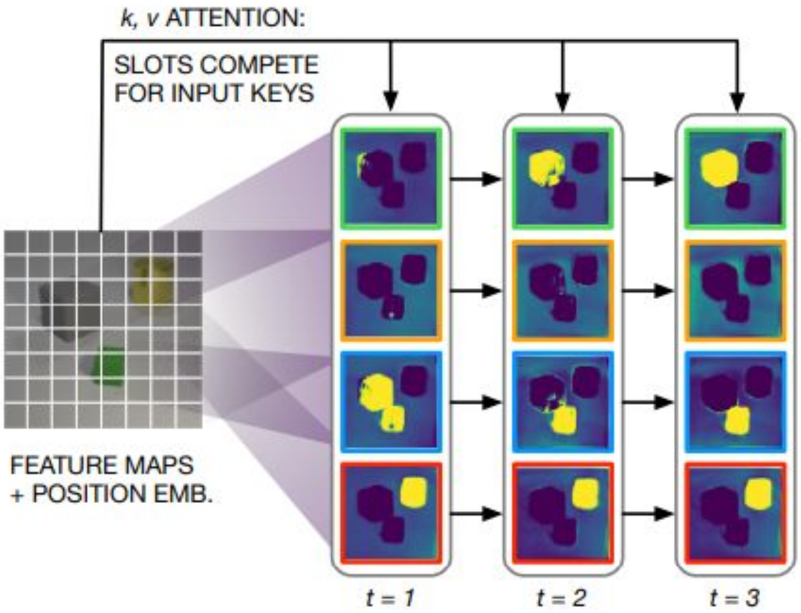
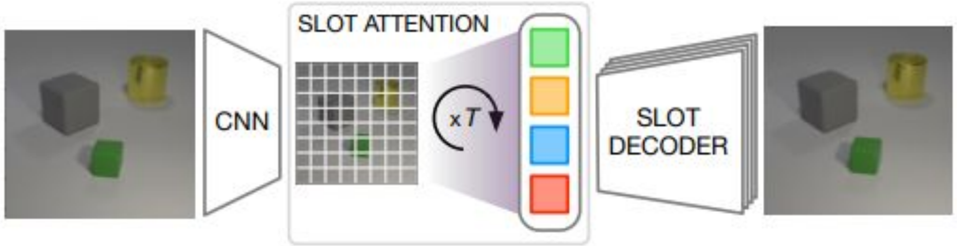


Image Source: https://www.hindawi.com/journals/complexity/2018/3927951/

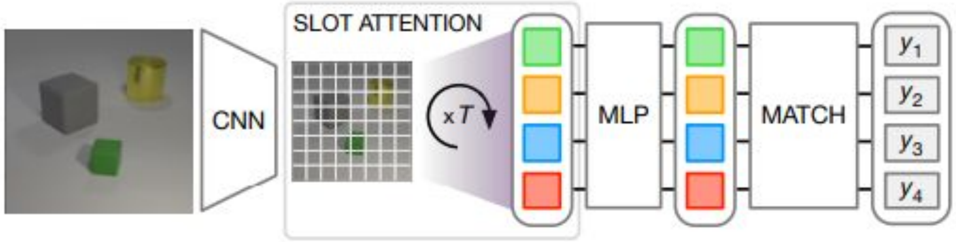
Example of Modularity: Slot Attention



(a) Slot Attention module.



(b) Object discovery architecture.



(c) Set prediction architecture.

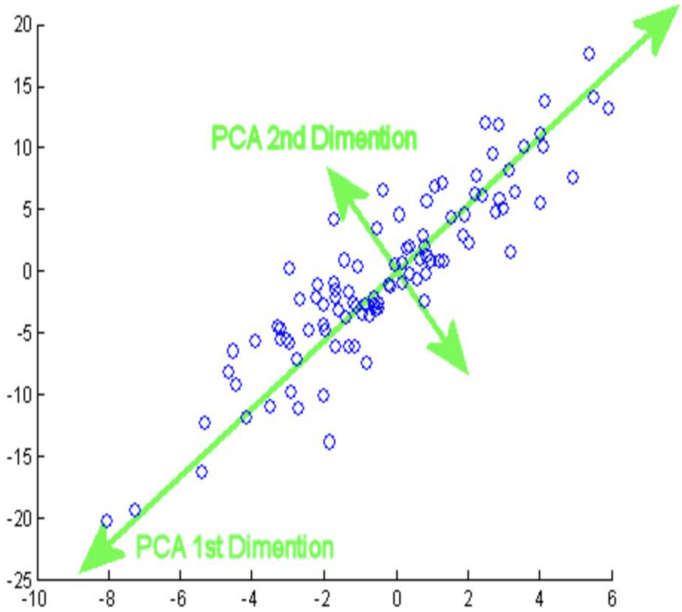
Feature Engineering

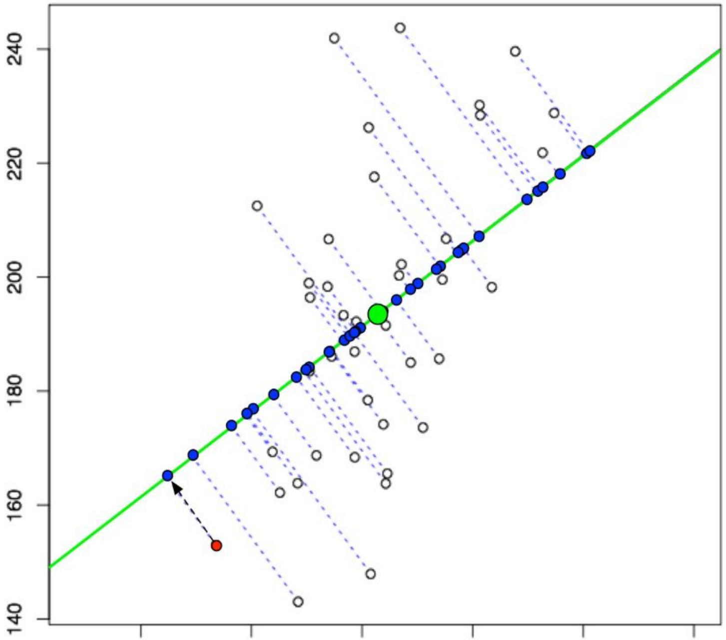
- Having more informative features makes the relationship that needs to be learned by the model simpler, allowing one to use other model-based interpretability methods.

- **Domain-Based Feature Engineering:** In each domain, expert knowledge and information obtained from data can be used to extract features (e.g. Use BMI instead of Mass and Height).

- **Model-Based Feature Engineering:** Automatic approaches to construct interpretable features.
  - Unsupervised methods: Clustering, Matrix factorization, Dictionary learning, Disentangled representation learning
  - Dimensionality Reduction: PCA, ICA, CCA

Example of Feature Engineering: PCA





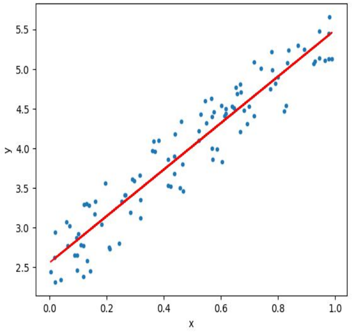
Post Hoc Interpretability

Post Hoc Interpretability

- Analyze a trained model to provide insight into the learned relationships
- Important when the collected data are high-dimensional and complex (e.g. images)
- Deals with the challenge that individual features may not be semantically meaningful
- **Two categories:**
    - Prediction-Level (Local): Explains individual predictions made by the models
    - Dataset- Level (Global): Focuses on global relationships the model has learned

Dataset-Level Interpretation

- **Feature Importance:**



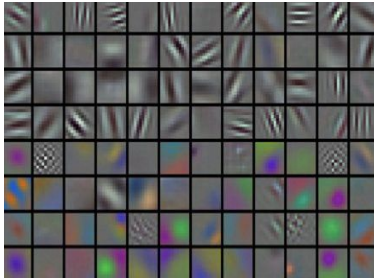
$$\hat{\beta} = \arg\min_{\beta_0,\ldots,\beta_p} \sum_{i=1}^{n} \left(y^{(i)} - \left(\beta_0 + \sum_{i=1}^{p} \beta_j x_j^{(i)} \right) \right)^2$$

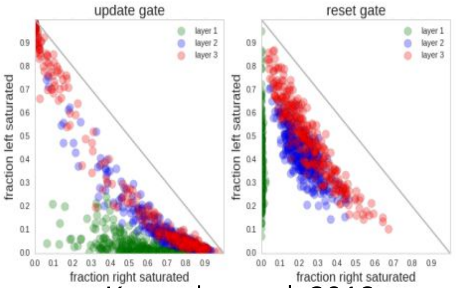
$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

- **Feature Interaction:**

	experience	degree	performance	sales	days_late
experience	36.922	19.184	3.1677	2.7169	0.90019
degree	19.184	44.714	2.3006	2.0608	0.4113
performance	3.1677	2.3006	30.562	12.954	1.1902
sales	2.7169	2.0608	12.954	30.19	1.3405
days_late	0.90019	0.4113	1.1902	1.3405	4.6451

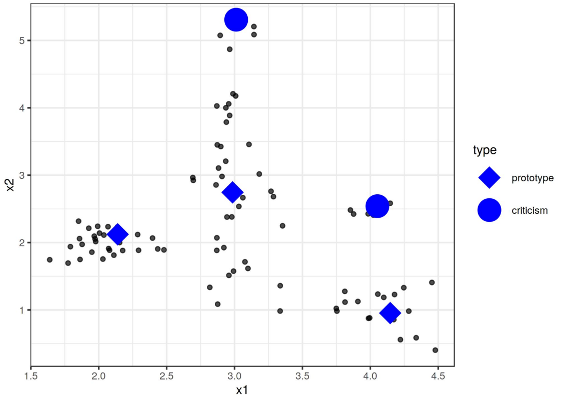
- **Visualization:**





Karpathy et al. 2018

- **Trends and Outliers:**

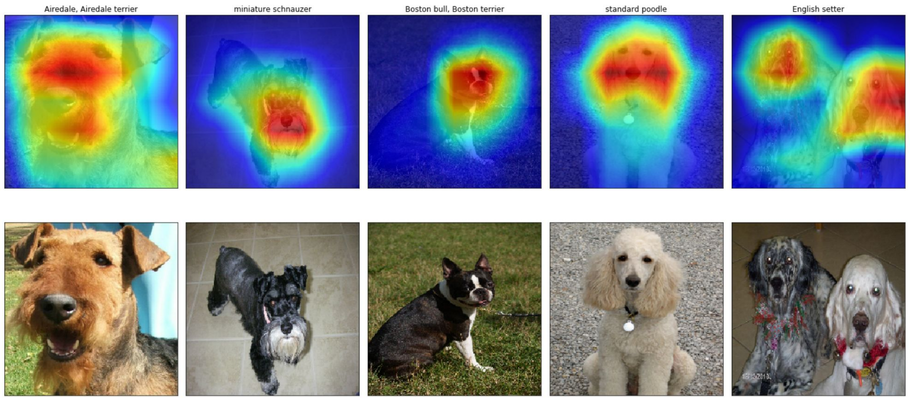


Prediction-Level Interpretation: Feature Importance

- Intuitively, a variable with a large positive (negative) score makes a highly positive or negative contribution to prediction for a particular instance.

- Can be useful for ensuring fairness of a model's decision.

- Unable to capture when algorithms learn interactions between variables. There are methods that evaluate relation between parameters and features.



BertViz

Future Work

Measuring Interpretation Desiderata

- **Measuring descriptive accuracy:**
  - Challenging to measure or quantify
  - An approach: Use Generative model to generate data, train a powerful model on the data, and see how interpretation method captures the relationships learned by the learned model.
  -
- **Demonstrating relevancy to real-world problems:**
  - **Common Pitfall**: focus on the novel output, ignoring what real-world problems it can actually solve.

  - **Domain-Specific Interpretations**
  - **Human Studies:** how much they trust a model's predictions

Model Based Interpretation

- **Usually fails to achieve a reasonable predictive accuracy**

- **Build accurate and interpretable models:** Devise new modeling methods which produce higher predictive accuracy while maintaining their high descriptive accuracy and relevance (e.g. Bayesian networks).

- **Tools for feature engineering:** The more informative the features, the simpler the model can be.
  - Improve unsupervised methods for feature engineering
  - Use visualisation, data exploration tools, interactive tools to enable the researchers to interact with and understand their data.

Post Hoc Interpretation

- Two Questions:
  - **What an interpretation of an ML model should look like?** There is a gap between the simple information provided by these interpretation methods and what the model has actually learned. Can we really close the gap? Should we consider a whole new framework?

  - **How post hoc interpretations can be used to increase a model's predictive accuracy?** post hoc interpretations uncover that a model has learned relationships a practitioner knows to be incorrect.


Conclusion

Conclusion

- There are multiple reasons that leads the researchers to model interpretability.
- There are different definitions of interpretability.
- PDR framework is proposed as desiderata for interpretability.
- Model-based and Post hoc methods have been proposed for model interpretability.
- However there are ambiguities surrounding desiderata of interpretability of a model and whether high predictive and descriptive accuracy can be achieved while fully understanding what deep models do.

References

1. Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd ed.). christophm.github.io/interpretable-ml-book/

Thanks!