21 - Interpreting deep models

Ludwig Krippahl

Interpretability

Summary

- Motivation
- Problems with black box models.
- Explanations and methods:

- Local Interpretable Model-agnostic Explanations (LIME)
- Layer-wise Relevance Propagation (LRP)
- Testing with Concept Activation Vectors (TCAV)
- Mapping concepts to ontologies

Motivation

Motivation for Explainable AI

- Problem 1: organize vacation photos

- Use DNN to classify photos with and without faces
- How does the network do it? Who cares...?

Motivation for Explainable AI

- Problem 1: organize vacation photos

- Use DNN to classify photos with and without faces
- How does the network do it? Who cares...?

- Problem 2: surgeon uses a DNN to recommend procedure

- Network processes radiological images and recommends extraction of left kidney
- Why? Without an explanation the surgeon cannot use this recommendation

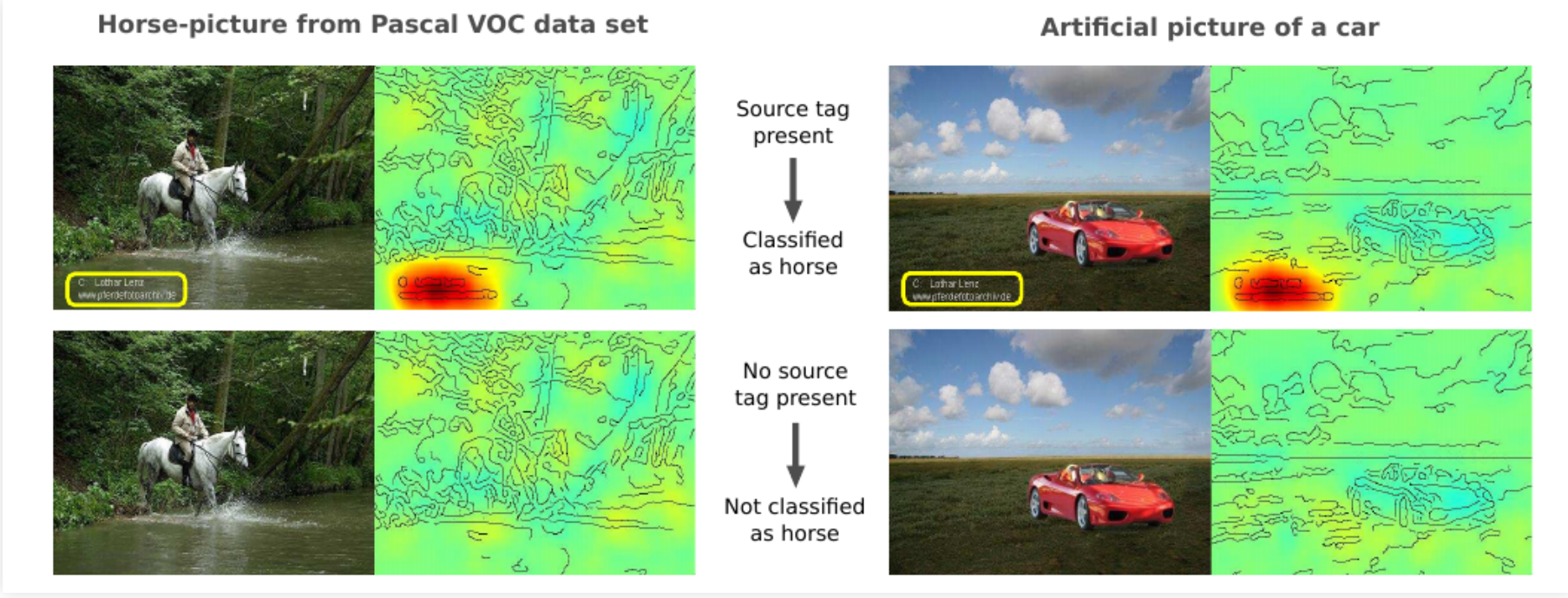
Motivation

Sometimes black box is not good enough

- Debugging or improving the system
- Trust that the system is working correctly
- Social acceptance of systems impacting our lives
- Ensure that a decision was reached correctly (and fairly)
- Auditing the system if something goes wrong
- For regulation, such as safety standards.
- For greater impact (e.g. automated recommendation)

Motivation

Identifying problems



Lapuschkin et al, Unmasking clever hans predictors, 2019

- The classifier was using the copyright tag to classify horses
- If added to a car image, it would be classified as a horse

Motivation

Trust and Accountability

- Without explanations it is easy to trust tools but not decisions

- AI is not only something we use but also something that decides for us

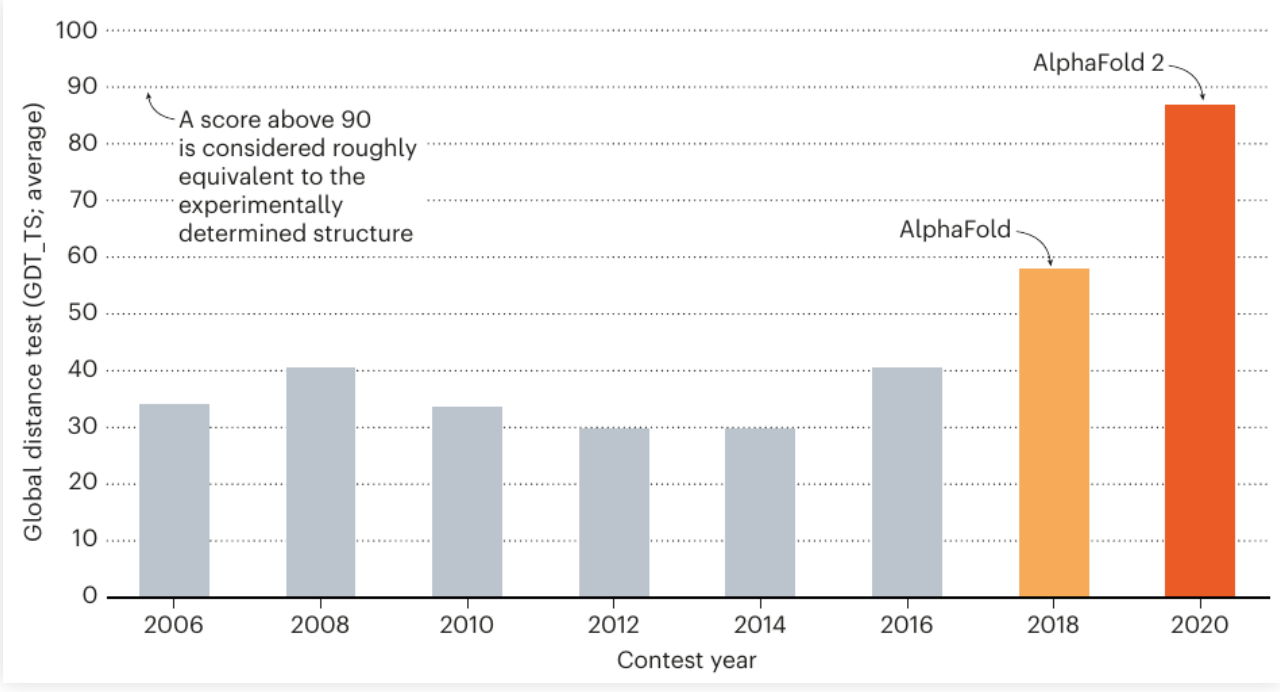
- Transparency is required for accountability and oversight.

- Regulation, for example

Motivation

Scientific Applications

- Deep learning has been very successful in scientific applications



Protein folding with AlphaFold2. Image from Callaway, 'It will change everything',2020

- But it would be great to understand how...

Motivation

Legal requirements

- EU's General Data Protection Regulation includes the right to an explanation
- Article 22:

" The data subject shall have the right not to be subject to a decision based solely on automated processing"

- The person deciding will need to understand the system's recommendation

Explanation

What is an explanation?

- Broad sense:

- A narrative that links different events and entities in an intelligible way
- Describes causal relations, consequences and the system explained.

- For our purpose, in practice, usefulness depends on target audience.

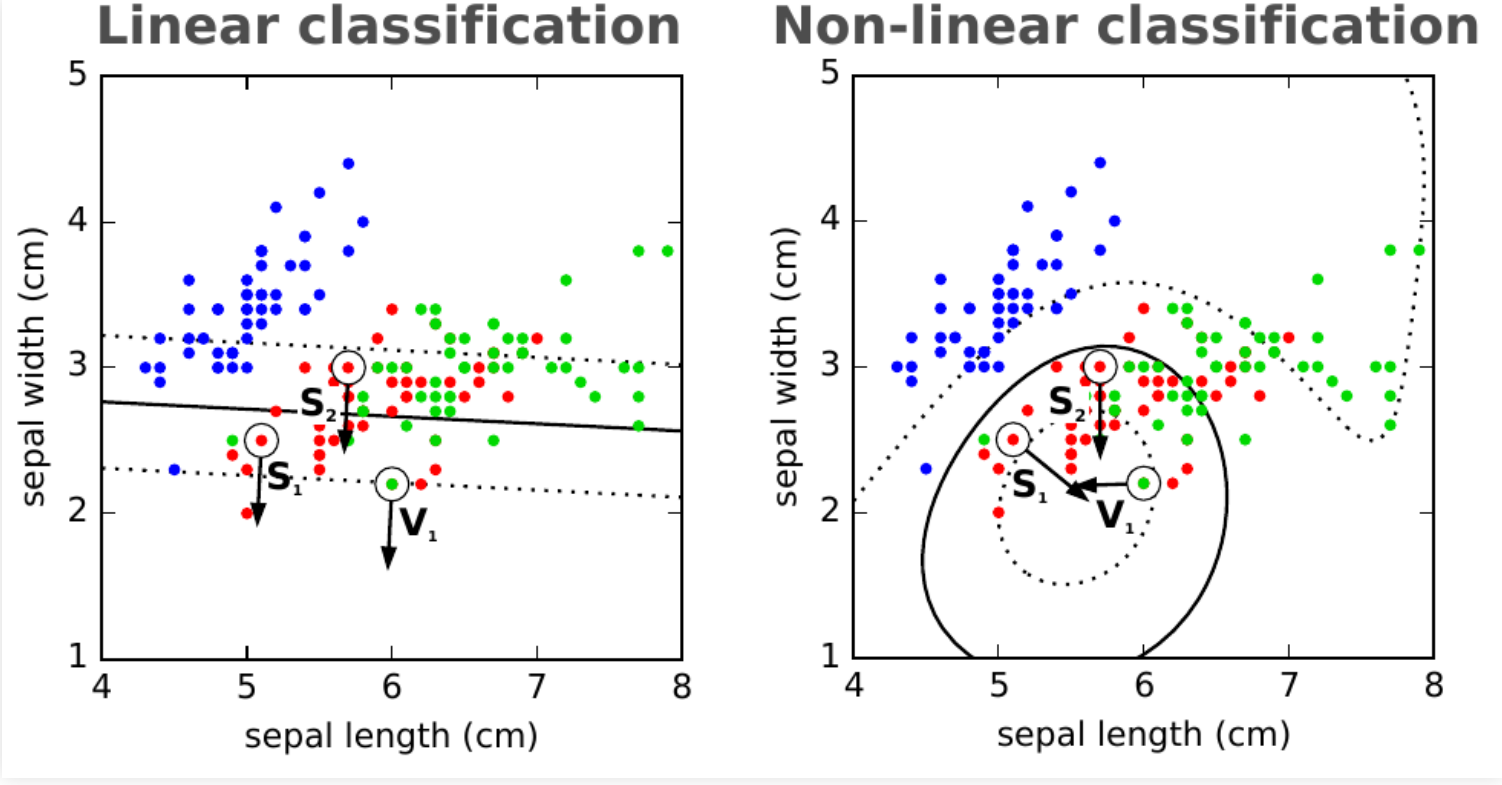
- The best explanation for the surgeon is not the best for the patient
- The developers of deep networks want explanations to help debug and optimize models
- The end user needs reasons to trust the output of the network.

- Choose interpretation methods for the target audience and purpose

LIME

Local Interpretable Model-agnostic Explanations

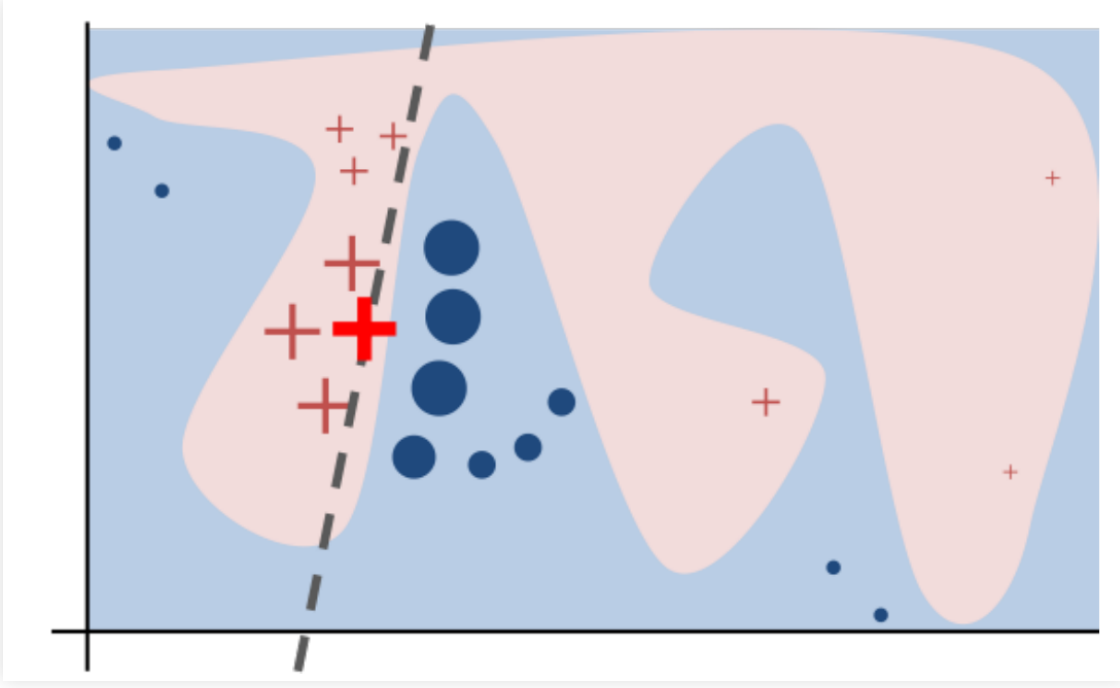
- Some simple models are easy to interpret. E.g. linear models



Lapuschkin et al, Unmasking clever hans predictors, 2019

Local Interpretable Model-agnostic Explanations

- Linear models are generally not powerful enough
- But can provide local approximations

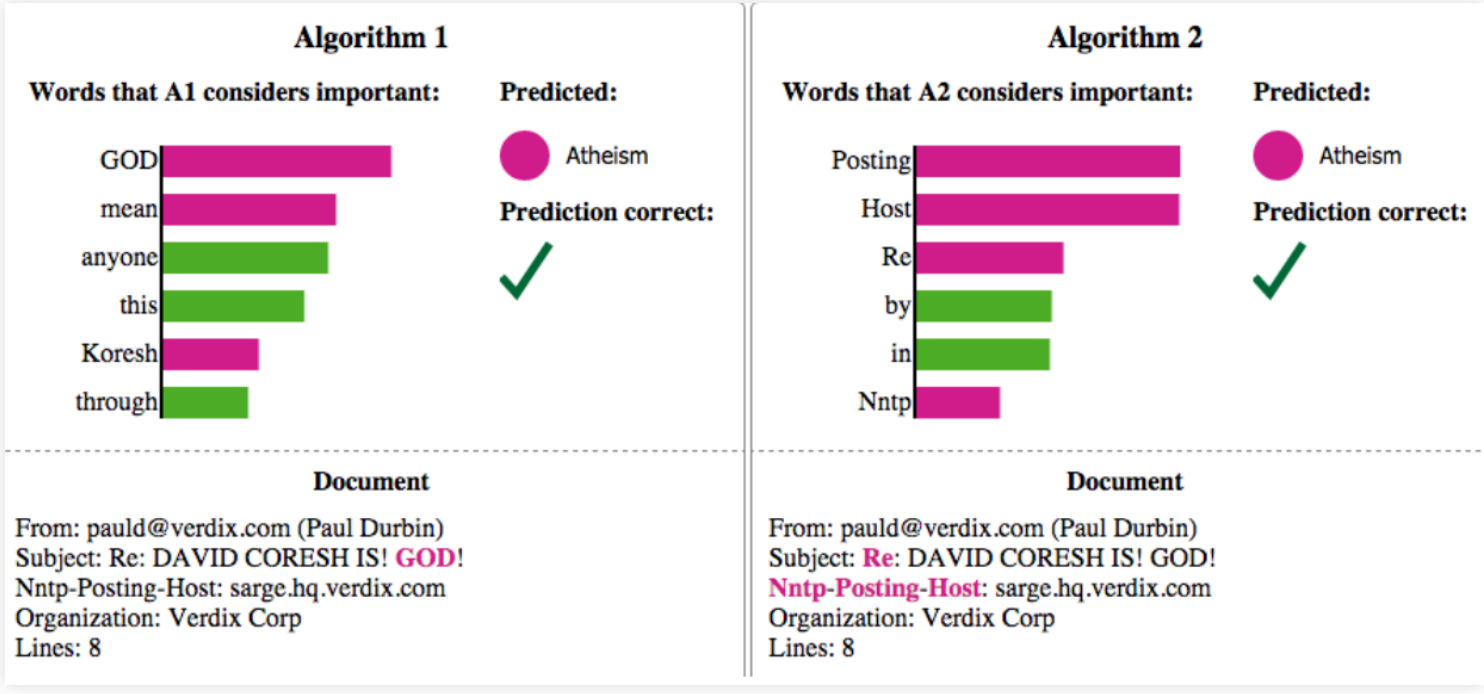


Ribeiro et al, Why should I trust you?,2016

Local Interpretable Model-agnostic Explanations

- Linear models are easy to interpret

- Both messages correctly classified as atheist, but with different features



Ribeiro et al, Why should I trust you?,2016

Local Interpretable Model-agnostic Explanations

- LIME does not provide a global explanation
- It provides a local explanation for a particular example
- It is model-agnostic because it does not care about how the model works
- It approximates the results with a linear classifier minimizing:

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} L(f, g, \pi_x) + \Omega(g)$$

- Where g is a linear model using any combination of features
- L measures the loss between explainer and model to be explained f
- $\Omega(g)$ is a measure of the complexity of model

LRP

Layer-wise Relevance Propagation

- Assign to each input a relevance measure for a particular output
- Takes into account the architecture and parameters of the trained model
- The relevance of the output neuron for a class is its activation
- Then propagate for neurons in preceding layers:

$$R_j = \sum_k \frac{a_j w_{jk}}{\sum_i a_i w_{i_k}} R_k$$

Layer-wise Relevance Propagation

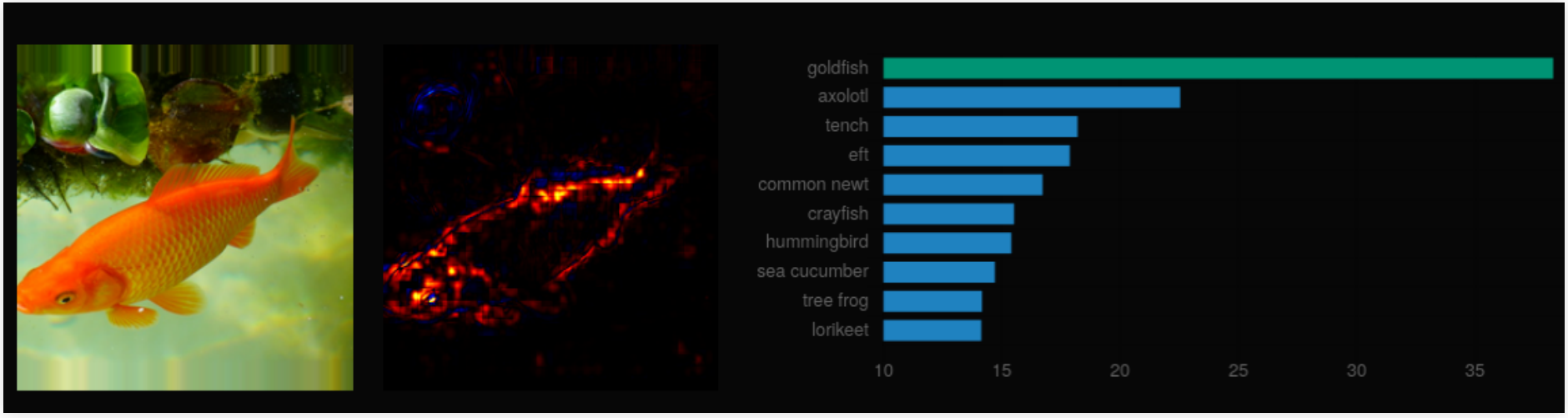
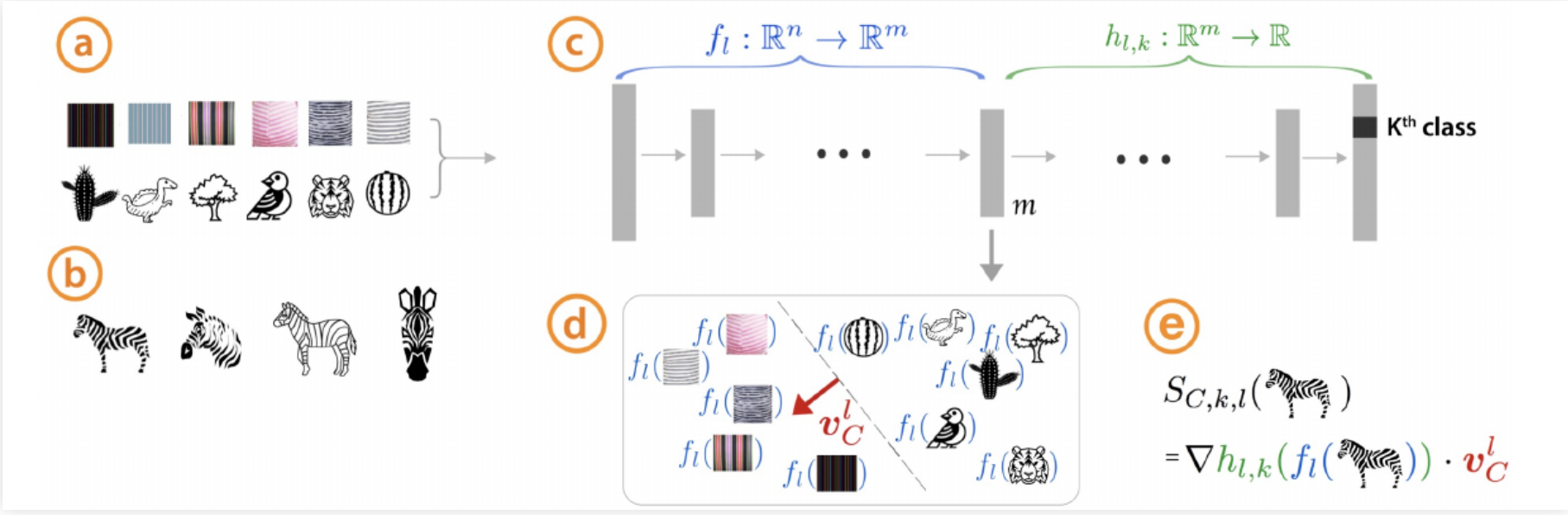


Image from Explainable AI demos at https://lrpserver.hhi.fraunhofer.de

TCAV

Testing with Concept Activation Vectors

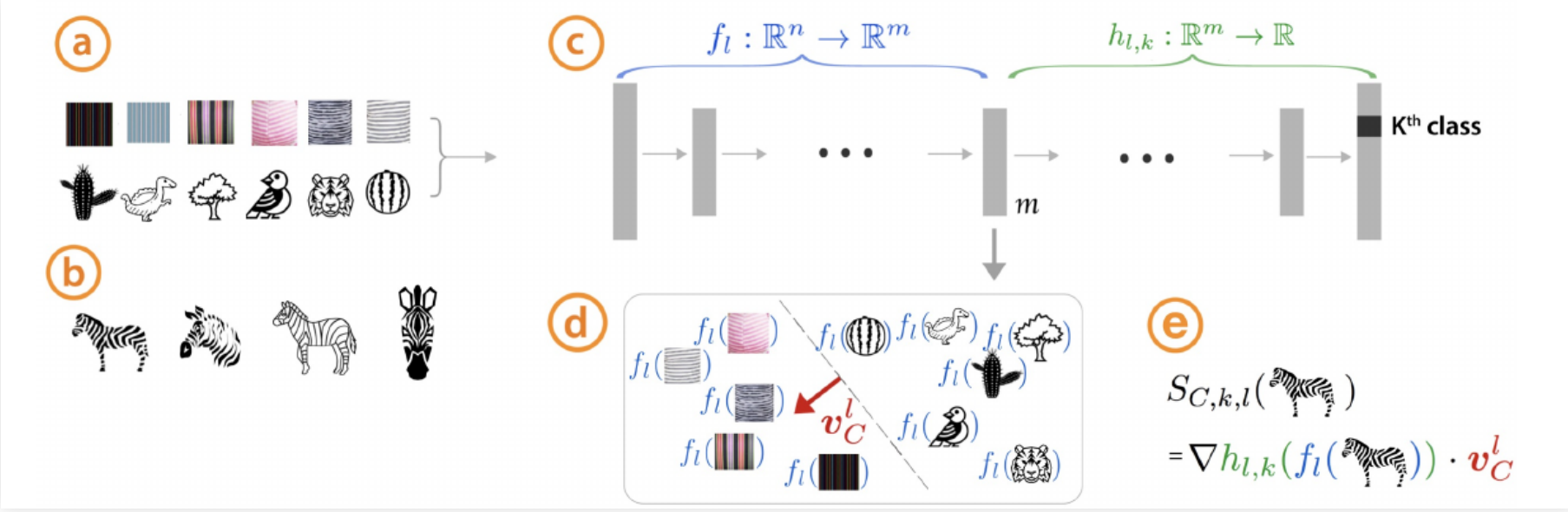
- Measure sensitivity of classifications to selected concepts



Kim et al, Interpretability beyond feature attribution, 2017

- a) Examples of a concept (striped) and random examples

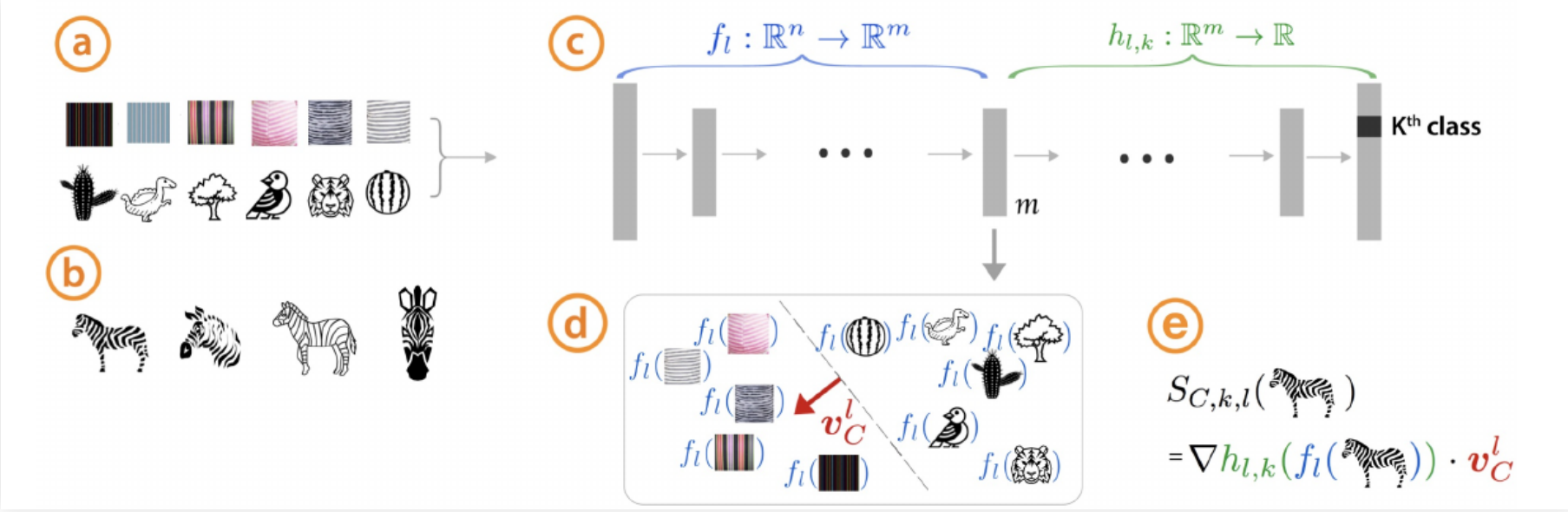
Testing with Concept Activation Vectors



Kim et al, Interpretability beyond feature attribution, 2017

- b) set of labelled examples of some class from the training data, such as zebras.

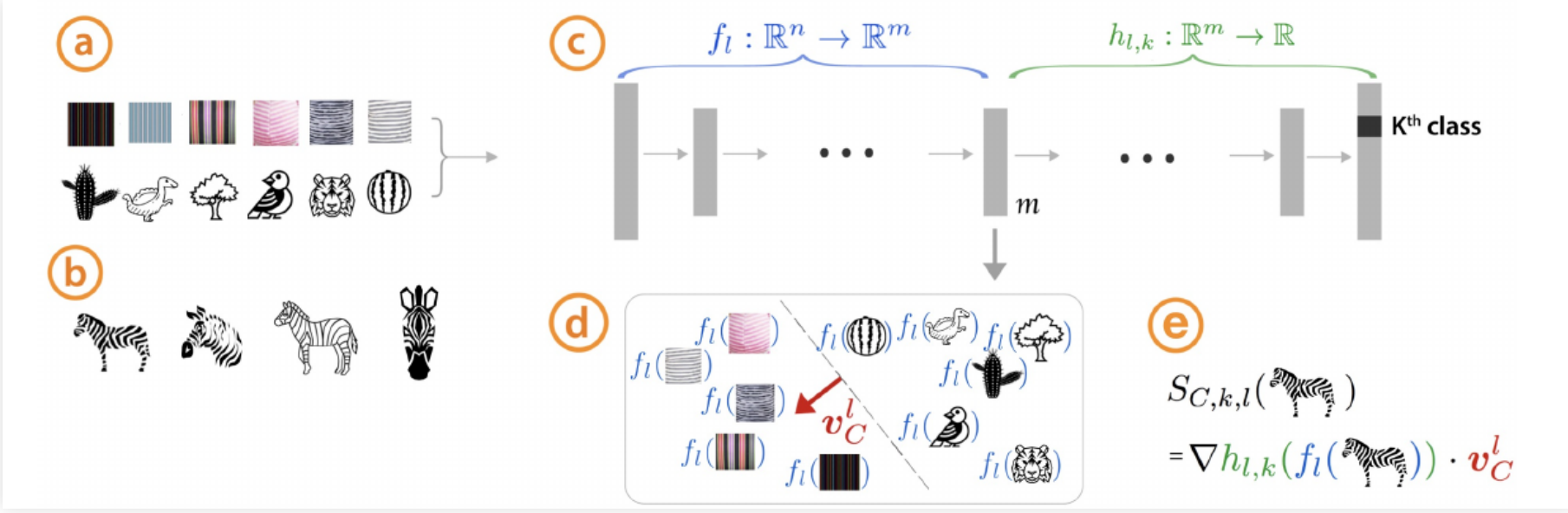
Testing with Concept Activation Vectors



Kim et al, Interpretability beyond feature attribution, 2017

- c) The trained classifier, includes the zebra class.

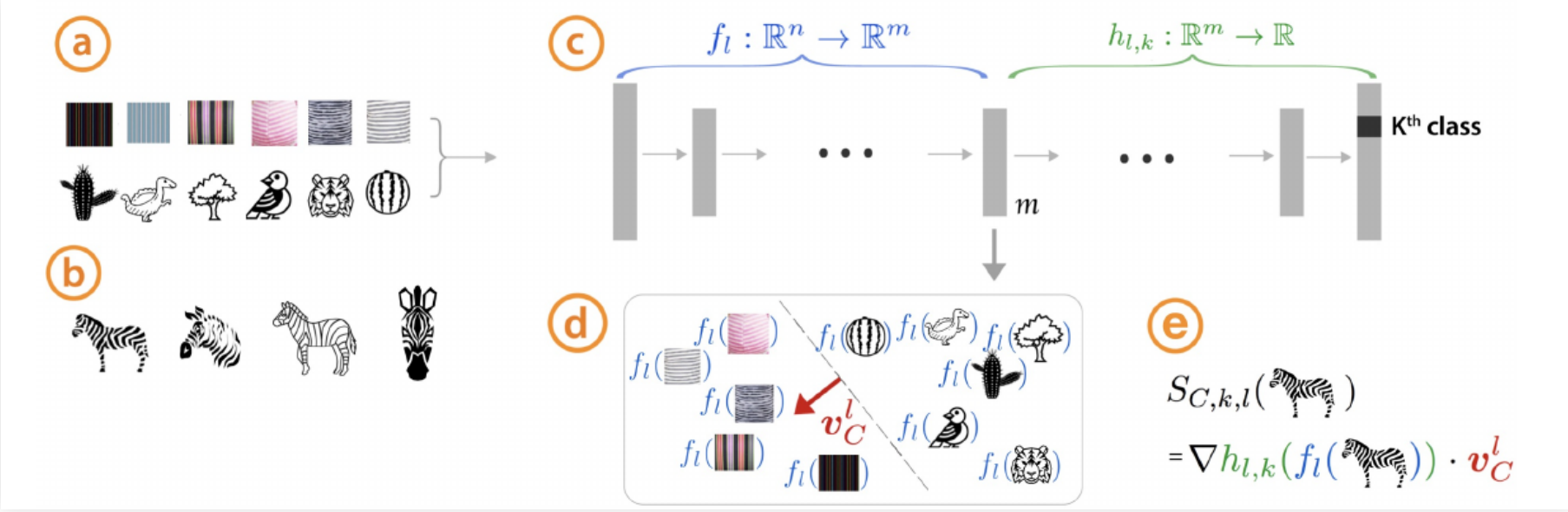
Testing with Concept Activation Vectors



Kim et al, Interpretability beyond feature attribution, 2017

- d) Vector normal to the linear decision surface at layer l (CAV)

Testing with Concept Activation Vectors

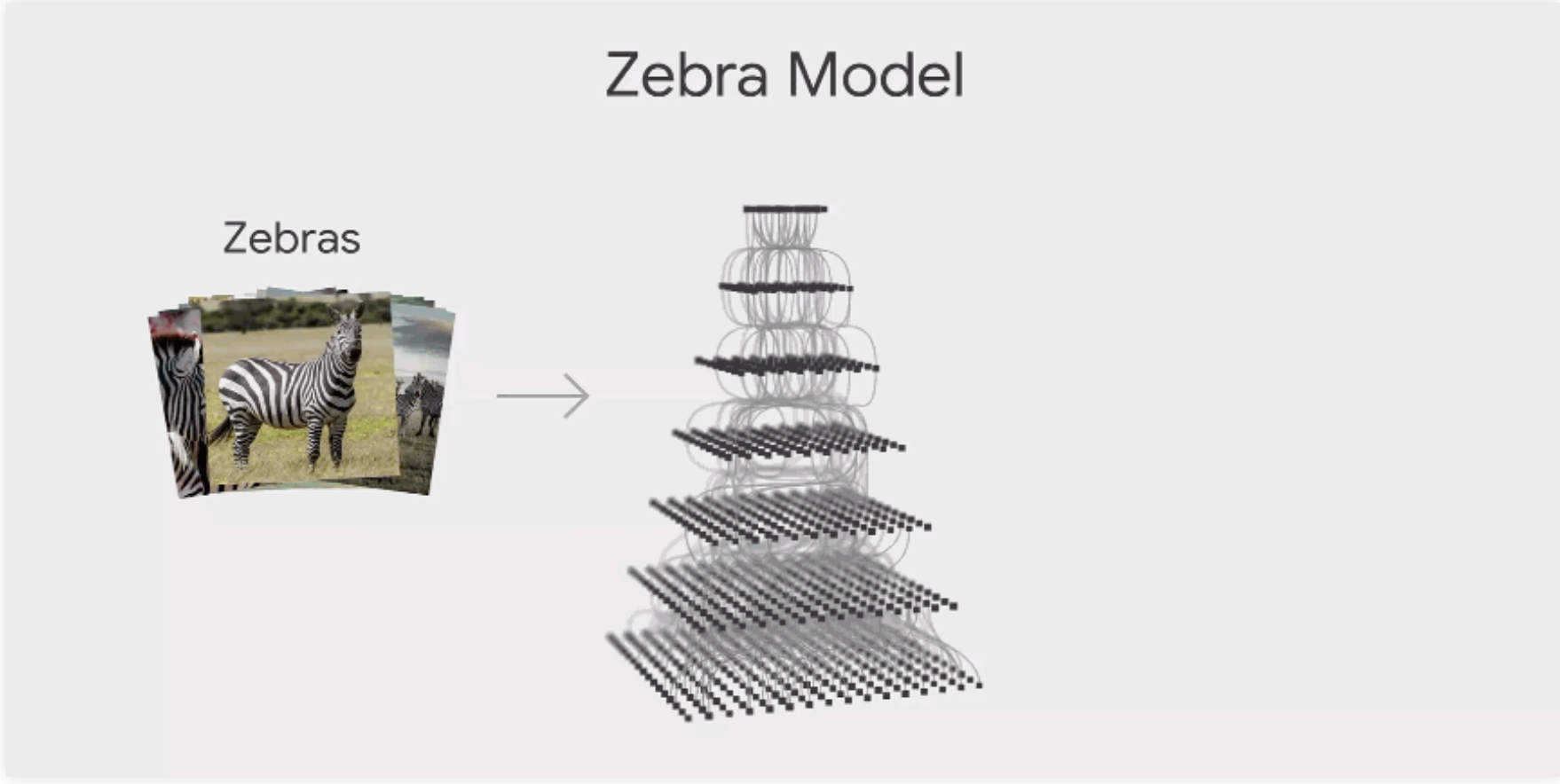


Kim et al, Interpretability beyond feature attribution, 2017

- e) Sensitivity of l to this concept, class and example is:

$$S_{C,k,l}(x) = \lim_{\epsilon \to 0} \frac{h_{l,k}\left(f_l(x) + \epsilon v_C^l\right) - h_{l,k}\left(f_l(x)\right)}{\epsilon}$$

Testing with Concept Activation Vectors



Google AI

Mapping concepts to ontologies

Mind the gap

Mapping concepts to ontologies

- (Work by Manuel Ribeiro, MIEI, 2020)
- Ontology:

- Formal specification of concepts and their logical relations.
- A STOP sign is an octagon, has a red background and STOP in white
- A warning sign is a triangle with a red border

- Goal: given a trained deep neural network, map activations to concepts

- Using examples illustrating different concepts in the ontology
- And auxiliary models, such as simple neural networks, for mapping

Mapping concepts to ontologies

- Advantage:

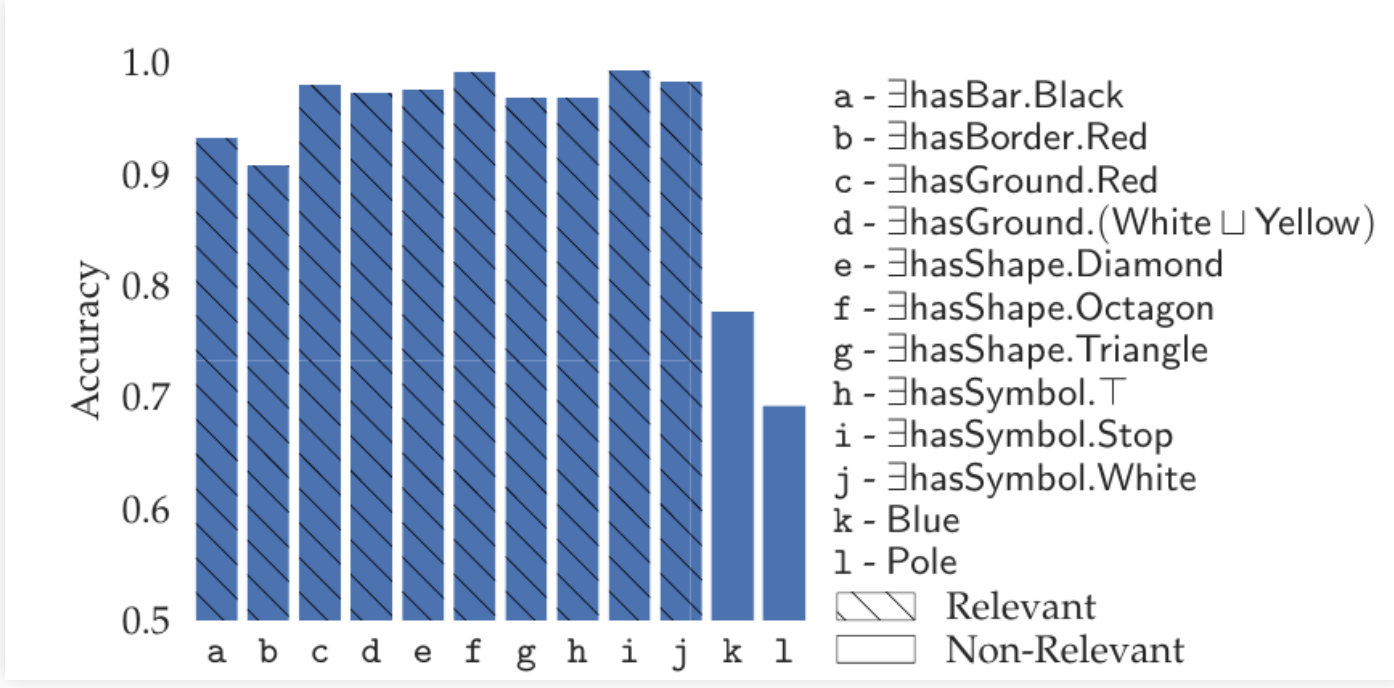
- After mapping we can use automated reasoning to generate justifications
- E.g. the network identified the concepts of red and octagon, which justifies concluding it is a STOP sign

Mind the gap

Mapping concepts to ontologies

- This also gives us insight into the manifold of the network representations

- Relevant concepts are easier to map as network learns to represent them



Risks of (partial) transparency

Risks

Transparency is desirable but...

- Interpretations that claim to make the model transparent create some risks
- Conflict of interests:

- My request for credit at the bank is denied
- I ask for an explanation
- The bank can use any of several interpretation methods to justify the decision

- The result may be a way to disguise unfair decisions
- In practice transparency must be a property of the whole process, not just applied to the model.

Summary

Interpretability

Summary

- Motivation:

- Transparency, trust, understanding

- Problems with black box models

- Debugging, auditing and regulation, responsibility

- Methods (examples):

- Local Interpretable Model-agnostic Explanations (LIME)
- Layer-wise Relevance Propagation (LRP)
- Testing with Concept Activation Vectors (TCAV)
- Mapping concepts to ontologies