XAI for Semantic Dependency

How to understand the impact of higher-level concepts on AI results

Holger Ziekow
Faculty of Business Information Systems
Furtwangen University
Germany
e-mail: holger.ziekow@hs-furtwangen.de

Peter Schanbacher
Faculty of Business Information Systems
Furtwangen University
Germany
e-mail: peter.schanbacher@hs-furtwangen.de
Presenter: Holger Ziekow

- Professor at Furtwangen University (Faculty of Business Information Systems)

- Research interest
  - Data Science and Machine Learning
  - XAI
  - Big and Streaming data
  - Application areas: (Manufacturing, Medicine, IoT, ...)

Hochschule Furtwangen
The General Problem

• Inner workings of *black box* machine learning models are hard to understand
• Humans seek insights into model decisions

Solution: XAI methods to analyze models
• XAI methods analyze black box models

• Our work introduces an XAI method to analyze how higher level concepts impact model decisions (semantic dependency)
Outline

• Background XAI and Partial Dependency Plots

• Our Extension: Semantic Partial Dependency Analysis (SDA)
  • Implementation with generators
  • Implementation with prediction models

• Experiments with Sample Implementation

• Conclusion and Future Work
XAI and Partial Dependency Plots

• Many methods exist to analyze the effect of individual features on a black box model
• Examples include SHAP and partial dependency plots (PDP)
Limitations of Partial Dependency Plots

- Analysis is limited to one feature at time\(^1\)
- Impact of higher level concepts (not directly reflected in a feature) cannot be visualized

\(^1\)or a small set of features

**Illustrative artificial example**

<table>
<thead>
<tr>
<th>Appetite</th>
<th>Pain</th>
<th>Attitude</th>
<th>Weight</th>
<th>Time to release</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1</td>
<td>3</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>1</td>
<td>12</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>5</td>
<td>14</td>
<td>11</td>
</tr>
</tbody>
</table>

Partial dependency on single feature

Partial dependency on higher level concept

Morbidity relates to predicted time to release
Another example for higher level concepts

- Consider an image classification task for landscapes
- Analyze how the presence of vegetation impacts the model outcome
  - Note: “presence of vegetation” is not a feature in the input data

![Diagram showing the probability of showing rural land (vs a city) vs the degree of presence of vegetation.](image-url)
SEMANTIC DEPENDENCY ANALYSIS
Semantic Dependency Analysis (SDA)

- Idea: compute the expected model output for data instances that have the higher level concept present to the defined degree $x_H$

- Generate samples from the modeled domain $X$ that have the concept to the specified degree $x_H$. (E.g. how much vegetation is present in a landscape image.)

$$SD_H(x_H) = E_X[\hat{f}(g(x_H, X))]$$

ML model

Random variable that returns feature vectors according to $x_H$ and $X$
Illustrative Example

Probability of showing rural land (vs a city)

Degree of presence of vegetation

- no trees
- few trees
- more than a few trees
- a lot of trees
Implementing Semantic Dependency Analysis

• How to implement $g(x_H, X)$?

• Proposed methods
  • Implementation with generators
  • Implementation with prediction models
Implementing $g$ with generators

Idea: Create synthetic data according to $X$ and ensure that the analyzed concept is present to degree $x_H$. 

$$g(x_H, X)$$

Data for analyzing the model with respect to $x_H$. 

Synthetic data generation

- Generative AI Models
- 3D-Engines
- Simulators
- ...

Hochschule Furtwangen
Implementing $g$ with prediction models

Idea: Use real data from distribution $X$ and filter out samples that have the analyzed concept is present to degree $x_H$.

$$SD_H(x_H, s) = \frac{1}{s-\epsilon} \int_0^s \left( 1 - \frac{d(x, x_H)}{s} \right) \, dx$$

Detection model (e.g. ML model)
Experiments with Sample Implementation
Experimental Setup

Image classification task as example

- Classify landscapes in “city” or “rural”
- Analyze the impact of the presence of trees on the model output
Generating Data for sample Classification Task

Using Stable Diffusion 2.0 with positive and negative prompts

**Class “city”**

**Positive prompt:** Photograph a city, high quality photography, Canon EOS R3

**Negative prompt:** digital art, drawing

**Class “rural landscape”**

**Positive prompt:** Photograph of a rural landscape, high quality photography, Canon EOS R3

**Negative prompt:** digital art, drawing
Generating Data for semantic dependency analysis

Class “cityNoTrees” (less than normal presence of trees)

Positive prompt: Photograph a city, high quality photography, Canon EOS R3
Negative prompt: digital art, drawing, trees
Class “cityTrees” (more than normal presence of trees)

Positive prompt: Photograph a city, trees, high quality photography, Canon EOS R3
Negative prompt: digital art, drawing
Class “TreesCity” (very high presence of trees)

Positive prompt: Photograph trees, city, high quality photography, Canon EOS R3
Negative prompt: digital art, drawing
Experimental Results\textsuperscript{1}

- SDA shows that the concept of “presence of trees” impact the classification in the expected way.
- The plausible result validates the viability of the approach.

\textsuperscript{1}more results in the paper
Conclusion

• **Contributions**
  - We demonstrated a way to analyze model dependency on higher level concepts
  - We described two general ways for implementation (through generators and detectors)
  - In experiments we demonstrate the feasibility in a sample implementation

• **Challenges**
  - Implementing generators and detectors with the desired behavior

• **Future work**
  - Exploring implementation of generators and detectors (e.g. 3D engines, diffusion models with image to image, etc.)