Enriching Visual with Verbal Explanations for Relational Concepts

Combining LIME with Aleph

Johannes Rabold, Hannah Deininger, Michael Siebers, Ute Schmid

29th of April, 2020

Cognitive Systems, University of Bamberg, Germany

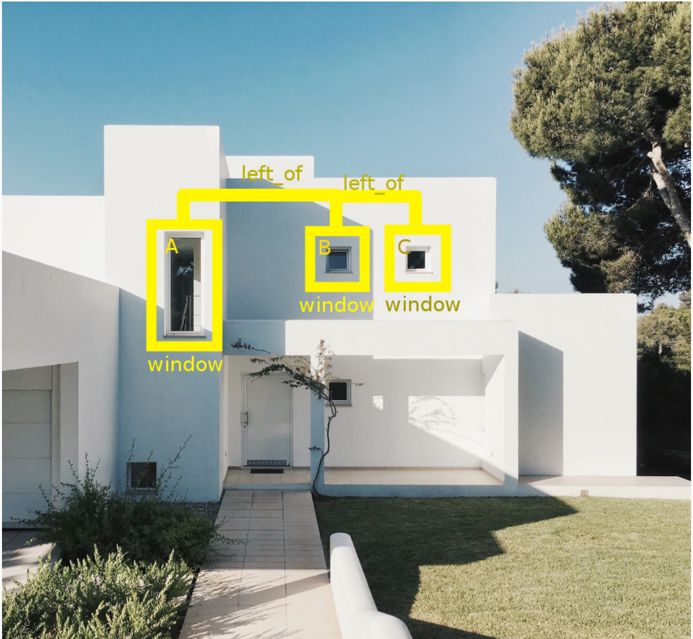
Towards Verbal Explanations

- Two approaches to interpretable models:
  - Build inherently interpretable models (Much overhead in preprocessing; bad performance)
  - **Extend good-performing black-box models by an interpretable component (Interpretability-Fidelity-Tradeoff)**

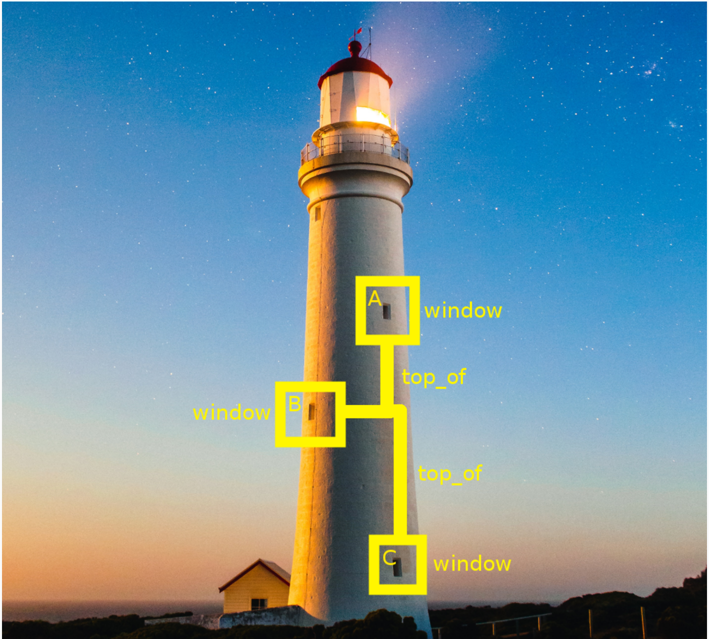
Towards Verbal Explanations

- Two approaches to interpretable models:
  - Build inherently interpretable models (Much overhead in preprocessing; bad performance)
  - **Extend good-performing black-box models by an interpretable component (Interpretability-Fidelity-Tradeoff)**
- Visualization as de facto standard for explanation of image classification decisions, but no possibility to highlight
  - Negations
  - **Feature values**
  - **Relations**

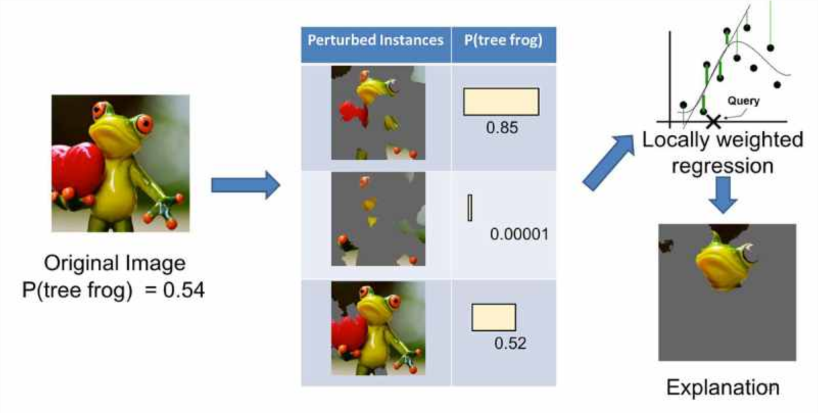
Illustrative Example Photo by Pixasquare uploaded to Unsplash



Illustrative Example Photo by Joshua Hibbert uploaded to Unsplash



Original LIME



Source: https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime

Inductive Logic Programming

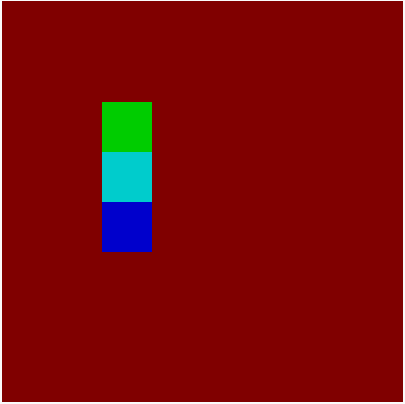
Inductive Logic Programming (ILP)

- Fits a logic theory over examples
- Input:
  - Positive and negative examples
  - Background Knowledge
- Output:
  - First Order Logic hypothesis

```
                    cat(C) :-
has_attribute(C, fur), has_attribute(C, whiskers).
```

- We used Aleph as a very flexible ILP framework

Background Information Extraction



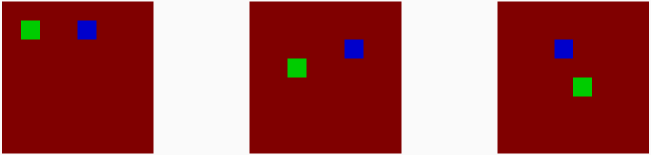
$\Longrightarrow$

`has_color(sp_18, green)`, `has_color(sp_26, cyan)`,
`has_color(sp_34, blue)`,
`on(sp_18, sp_26)`, `on(sp_26, sp_34)`

Simplified algorithm

```
Require: Instance $x \in X$
Require: Classifier $f$, Selection size $k$, Threshold $\theta$
    $S \leftarrow LIME(f, x, k)$
    $A \leftarrow$ extract_attribute_values($S$)
    $R \leftarrow$ extract_relations($S$)
    for each $r(i, j) \in R$ do
        $z \leftarrow flip\_in\_image(x, i, j)$
        $r' \leftarrow r(j, i)$
        $R' \leftarrow R \backslash \{r\} \cup \{r'\}$
        $R' \leftarrow calculate\_side\_effects(R', r')$
        if $f(z) \geq \theta$ do
            $E^+ \leftarrow E^+ \cup \{\langle A, R' \rangle\}$
        else
            $E^- \leftarrow E^- \cup \{\langle A, R' \rangle\}$
    end for
    $T \leftarrow Aleph(E^+, E^-)$
return $T$
```

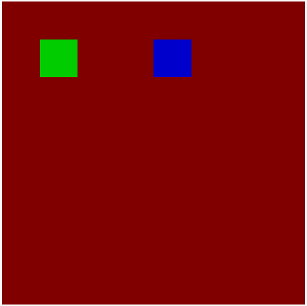
Experiment 1 - Single Relation Concept



(a) (b) (c)

Positive (a, b) and negative (c) examples for the concept “Green left of Blue“.

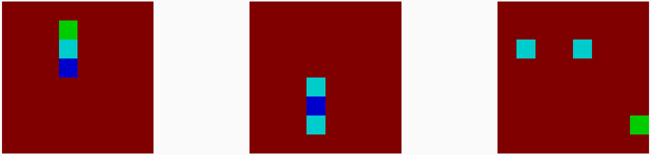
Experiment 1 - Single Relation Concept



$k = 3$, $\theta = 0.8$

```
concept(A) :- contains(B, A), has_color(B, green),
contains(C, A), has_color(C, blue), left_of(B, C).
```

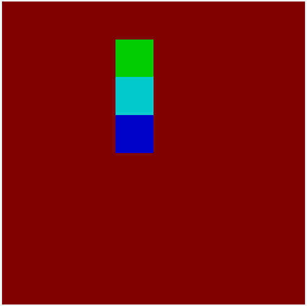
Experiment 2 - Concept “Tower“



(d) (e) (f)

Positive (a) and negative (b, c) examples for the concept “tower“.

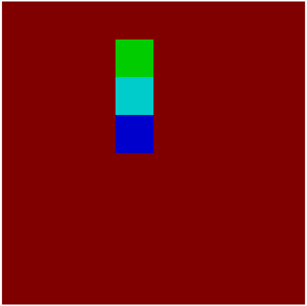
Experiment 2 - Concept “Tower“



$k = 3$, $\theta = 0.8$

```
concept(A) :- contains(B, A), has_color(B, cyan),
              contains(C, A), on(B, C).
```

Experiment 2 - Concept “Tower“



$k = \mathbf{4}$, $\theta = 0.8$

```
concept(A) :- contains(B, A), has_color(B, cyan),
contains(C, A), has_color(C, blue), top_of(B, C).
```

Future Work

- Evaluating interpretability of explantions in a user study
- Using real-world datasets

$\Longrightarrow$ Extracting semantically interpretable features from images

- Use interactive learning approaches for semi-automated annotation
- Use method to explain Capsule Networks

Thank you for your attention!

Questions?

`johannes.rabold@uni-bamberg.de`