Finding Shortcuts to a black-box model using Frequent Sequence Mining

Can Frequent Sequence Mining help find short-cuts for a complex black-box model?

BACKGROUND

Deep-learning (DL) models are reffered to as hard-to-explain. Various techniques have been proposed to use local explanations for the behaviour of DL models, but little attention has been paid to global explanations.

Frequent sequence mining generalizes connections between a model's input and output, generating rules to global explanations for the model.

Our research question: *can frequent sequence mining find short-cuts to a complex black-box model?*

METHODOLOGY

The main approach is to make shortcuts for
- state-of-the-art prediction model ExPred [1],
- which is trained on FEVER [2] for fact-checking,
- using DESQ [3] as a Frequent Sequence Mining tool

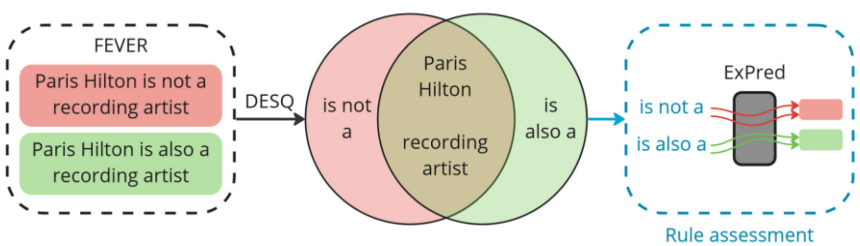

Figure 1: rule mining process on an example claim from FEVER

Main metrics for the assessment of rules:
- Support, a measure of the coverage of a rule.
  $$Supp(A \rightarrow B) = P(A \cup B)$$
- Confidence, a measure of the strength of a rule.
  $$Conf(A \rightarrow B) = P(B|A)$$
- Attack success rate, a measure of the succes of using rules in adversarial prompts for attacking the model.
  $$Success = \frac{successful\ examples}{total\ examples}$$

RESULTS

The patterns found in FEVER were visualised into the three categories of Figure 2. As shown, adverbs and adjectives can be short-cuts to the model refuting a claim. Conversely, existential clauses make the model support a claim. Most sequence patterns are found in the neutral set, which reveal the focus of the training dataset, as well as trivial language building blocks as illustrated in Table 1.

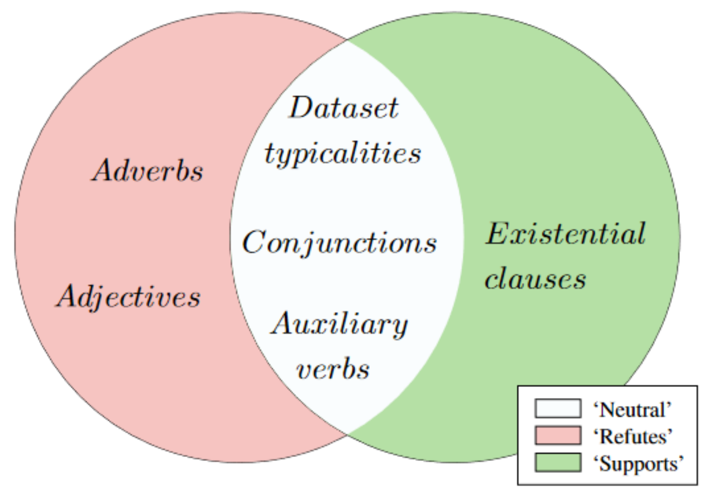

Figure 2: Venn-diagram depicting part-of-speech categories of patterns in FEVER[2].

The strongest rules in FEVER succesfully create shortcuts for the ExPred model as seen in Table 2.

$s \in$ 'Refutes'		$s \in$ 'Neutral'		$s \in$ 'Supports'	
s	$Supp(s)$	s	$Supp(s)$	s	$Supp(s)$
refused	0.36%	and	78%	acted	0.67%
yet	0.35%	the	70%	contains	0.29%
exclusively	0.31%	is	58%	birth	0.29%
unable	0.19%	a	57%	helped	0.05%

Table 1: selection from each class of the four most frequent single-item sequence patterns in FEVER[2].

DISCUSSION

Our results expose potential vulnerabilities in ExPred, and we show how the rules can be used for risk assessment. However, since the adversarial prompts were manually forged, the success-rates might be higher using automation.

s	$\rightarrow r(s)$	FEVER	ExPred	$Success(\bar{s})$
is incapable of being	→ Refutes	100%	94%	78%
has only ever been	→ Refutes	100%	99%	62%
does not have	→ Refutes	100%	85%	83%
is exclusively	→ Refutes	100%	99%	60%
is not a(n)	→ Refutes	100%	100%	74%
has yet to	→ Refutes	100%	100%	90%
is only a(n)	→ Refutes	100%	99%	77%
was unable to	→ Refutes	100%	95%	76%
was incapable of	→ Refutes	100%	97%	89%
There is a	→ Supports	100%	90%	89%

Table 2: selection of the 10 strongest rules and their success as adversarial attacks to the model.

CONCLUSIONS

Main findings:
- The ExPred model relies on shortcuts when making predictions.
- The rules can be a risk assessment tool for DL models using counterfactual attacks.

Future work:
- Assessment of a larger population of shortcuts
- Application to other datasets and models
- Extend and automate adversarial prompt attacks

REFERENCES

[1] Zijian Zhang, Koustav Rudra, and Avishek Anand. (2021) "Explain and Predict, and then Predict Again".
[2] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. (2018). "FEVER: a large-scale dataset for Fact Extraction and VERification".
[3] Kaustubh Beedkar and Rainer Gemulla. (2016). "DESQ: Frequent Sequence Mining with Subsequence Constraints".

Author Jean-Paul Smit **Contact** j.p.smit@student.tudelft.nl

TU Delft Delft University of Technology

Supervisors Lijun Lyu, Lorenzo Corti **Responsible Professor** Avishek Anand