Towards Robust Interpretability with Self-Explaining Neural Networks - Reproduction

Rico Mossinkoff
Yke Rusticus
Roberto Schiavone
Ewoud Vermeij

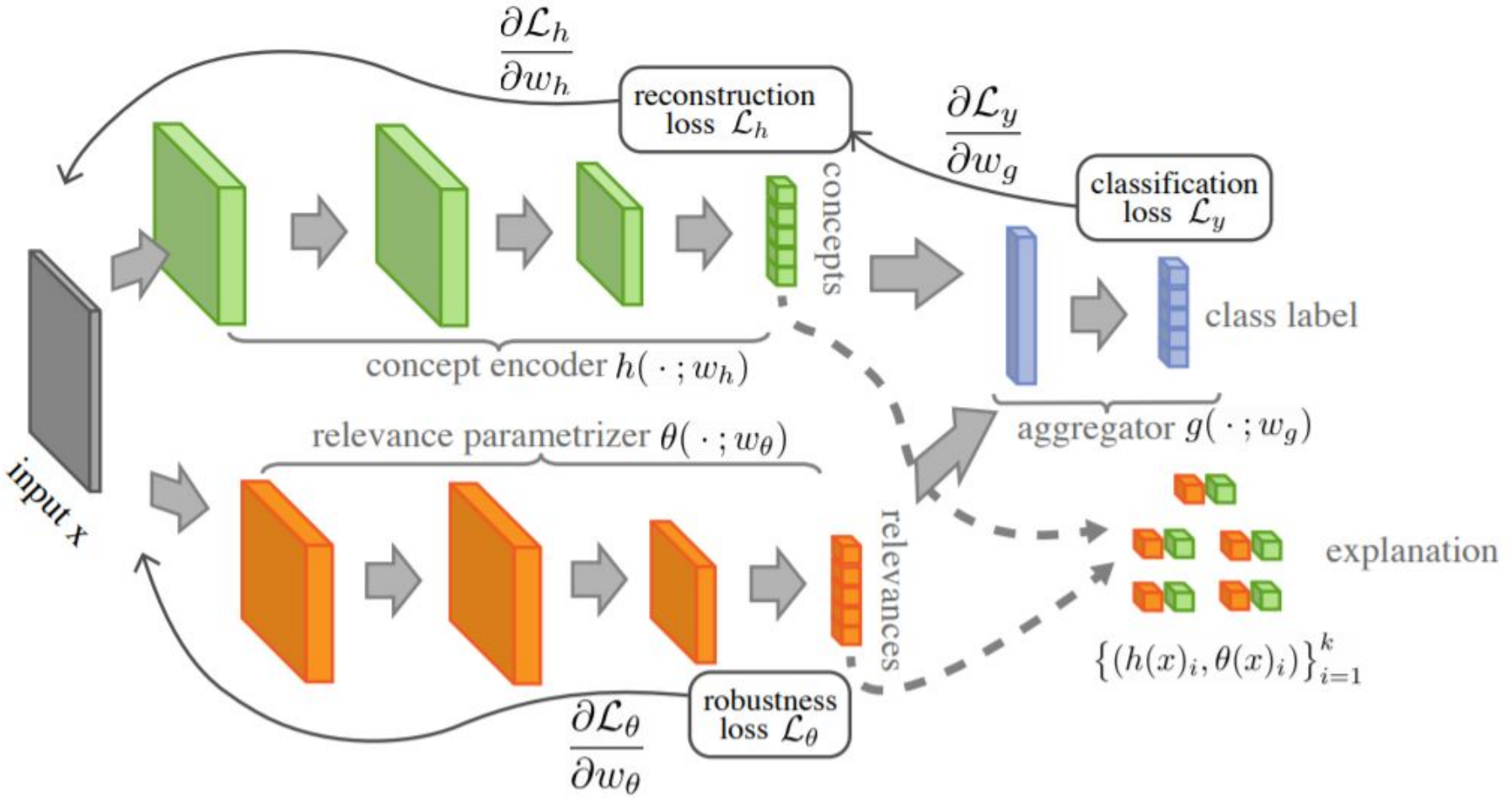
Table of contents

Introduction

- AI becomes black box
- Fairness and transparency is hard
- Alvarez-Melis and Jaakkola (SENN)
- Extended with different visualizations

Self-Explaining Neural Network (SENN)

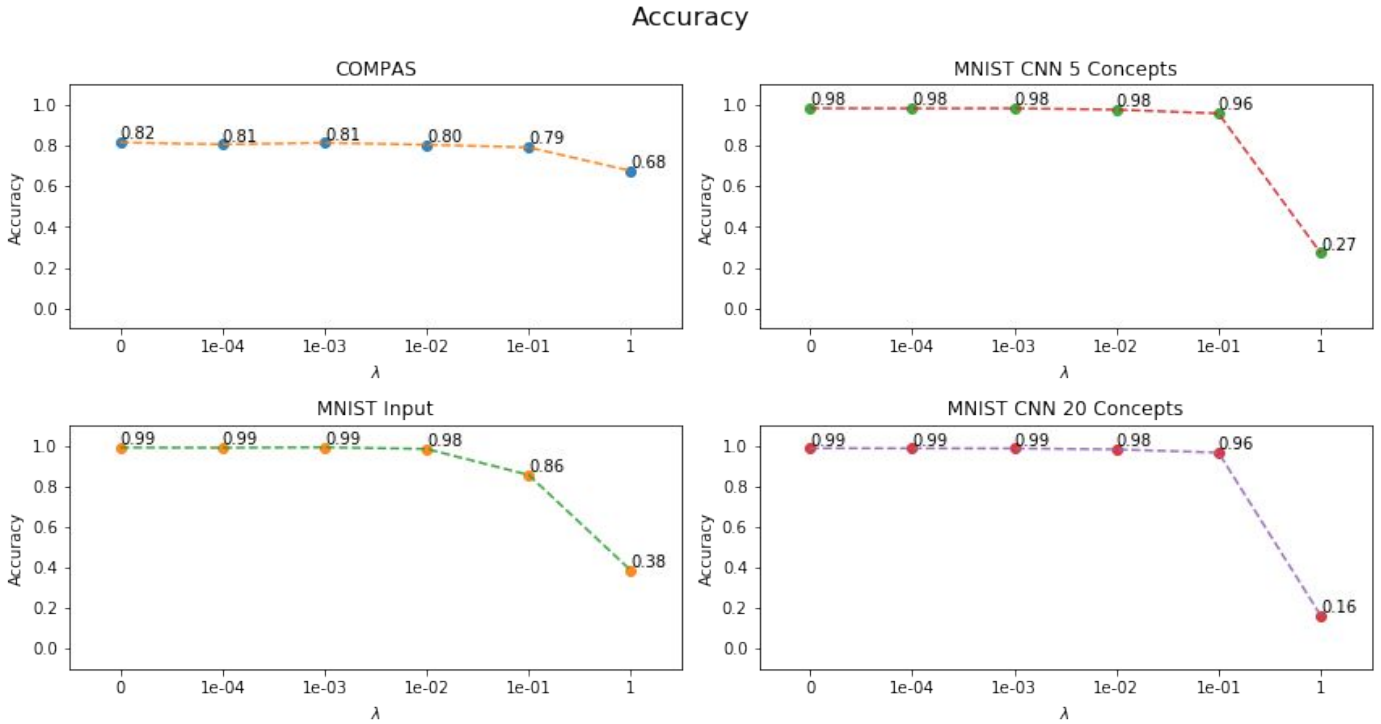
- SENN (Self Explaining Neural Network)
  - Basis concepts
  - Parametrizer
- Basis of a SENN:
  - Linear model $f(x) = \theta x$
- Complex SENN:
  - $f(x) = \theta(x)\, x$
  - $f(x) = \theta(x)\, h(x)$


$\frac{\partial \mathcal{L}_h}{\partial w_h}$
reconstruction loss $\mathcal{L}_h$
$\frac{\partial \mathcal{L}_y}{\partial w_g}$
classification loss $\mathcal{L}_y$
concepts
class label
concept encoder $h(\cdot\,; w_h)$
aggregator $g(\cdot\,; w_g)$
relevance parametrizer $\theta(\cdot\,; w_\theta)$
input x
relevances
explanation
$\{(h(x)_i, \theta(x)_i)\}_{i=1}^{k}$
robustness loss $\mathcal{L}_\theta$
$\frac{\partial \mathcal{L}_\theta}{\partial w_\theta}$

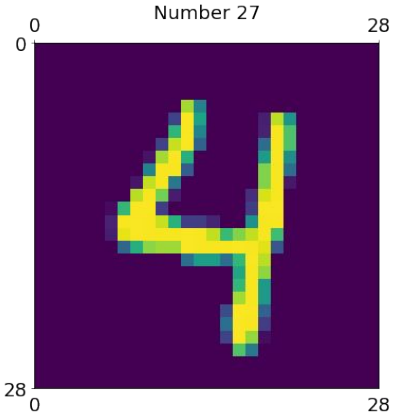
Methods and experimental setup

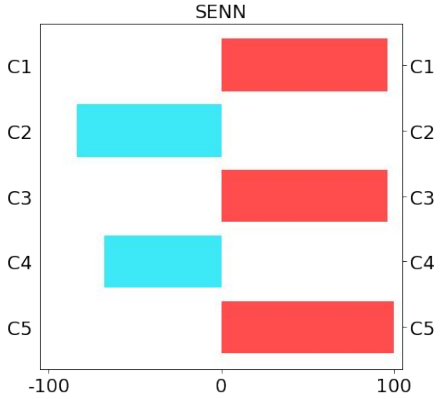
- 3 desiderata for basis concepts
  - Fidelity (relevance)
  - Diversity (representable inputs)
  - Grounding (human understandable)
- 3 evaluation criteria on explainability
  - Explicitness/Intelligibility (how understandable are SENN explanations)
  - Faithfulness (are relevant features truly relevant)
  - Stability (local linearity is needed to keep the model explainable)
- Datasets
  - COMPAS
  - MNIST

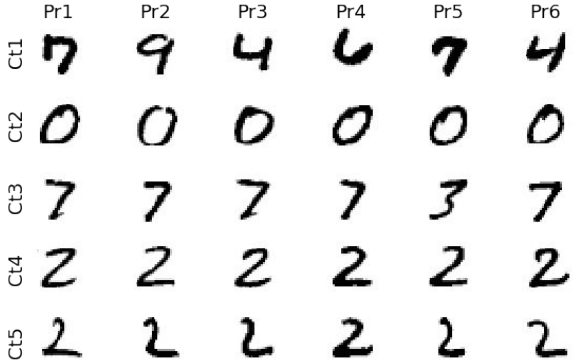
Results - Reproducibility



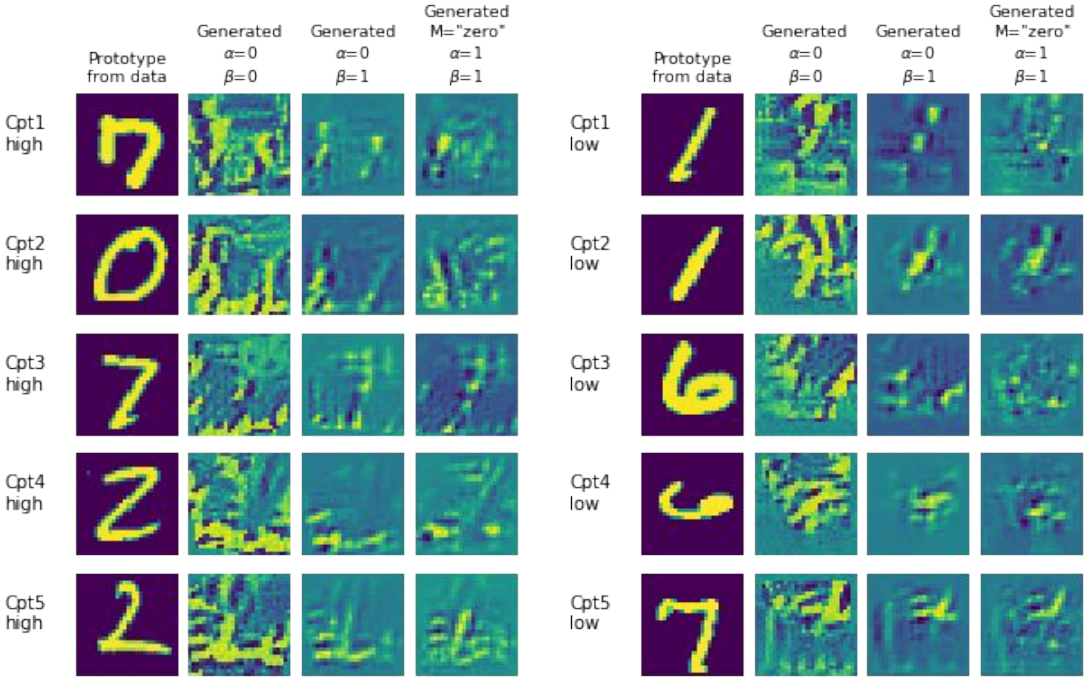
Results - Explicitness/Intelligibility



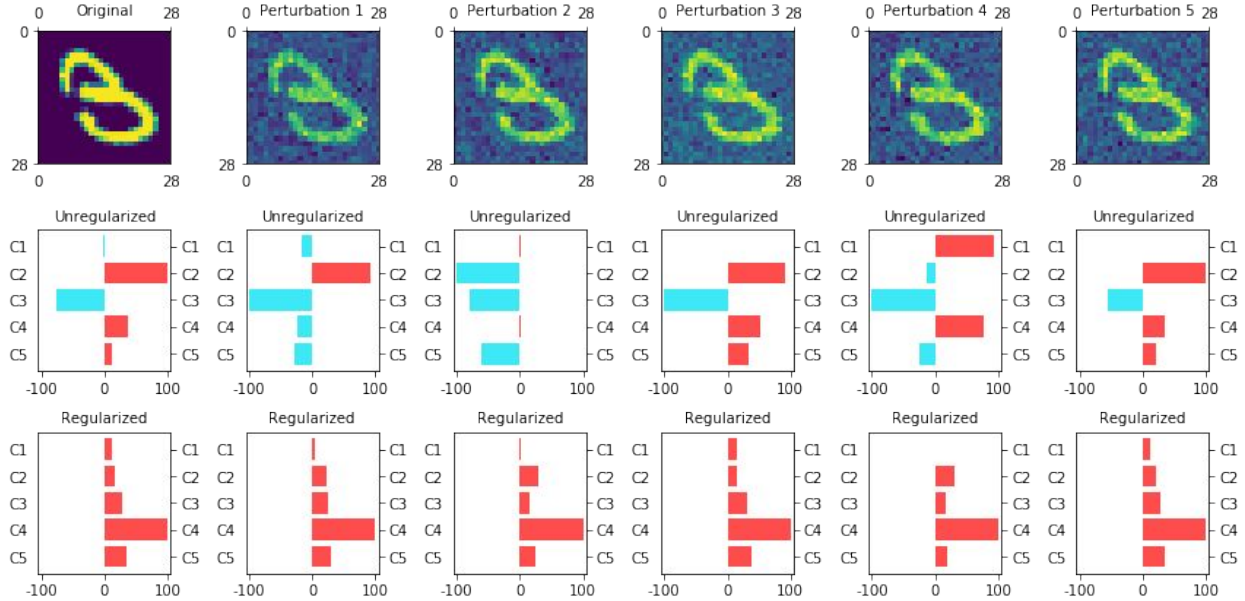




Results - Explicitness/Intelligibility



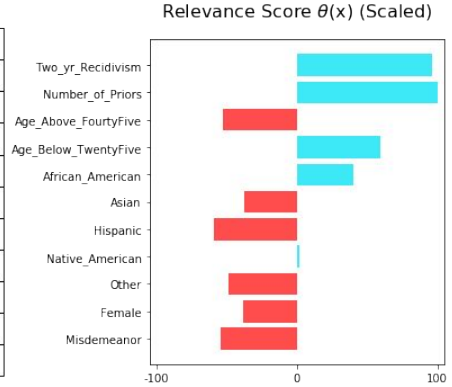
Results - Stability



Results - Stability

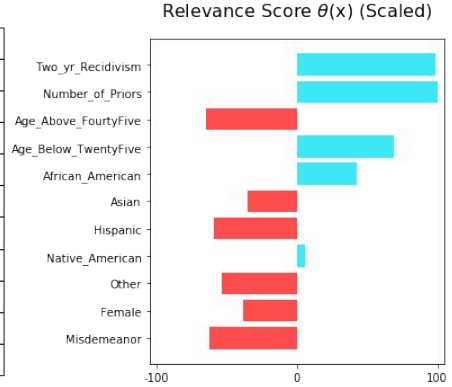
Input Value

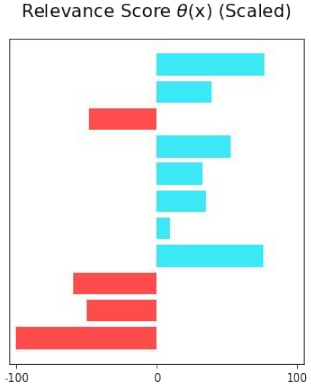
Two_yr_Recidivism	0.0
Number_of_Priors	0.23
Age_Above_FourtyFive	0.0
Age_Below_TwentyFive	1.0
African_American	1.0
Asian	0.0
Hispanic	0.0
Native_American	0.0
Other	0.0
Female	0.0
Misdemeanor	0.0



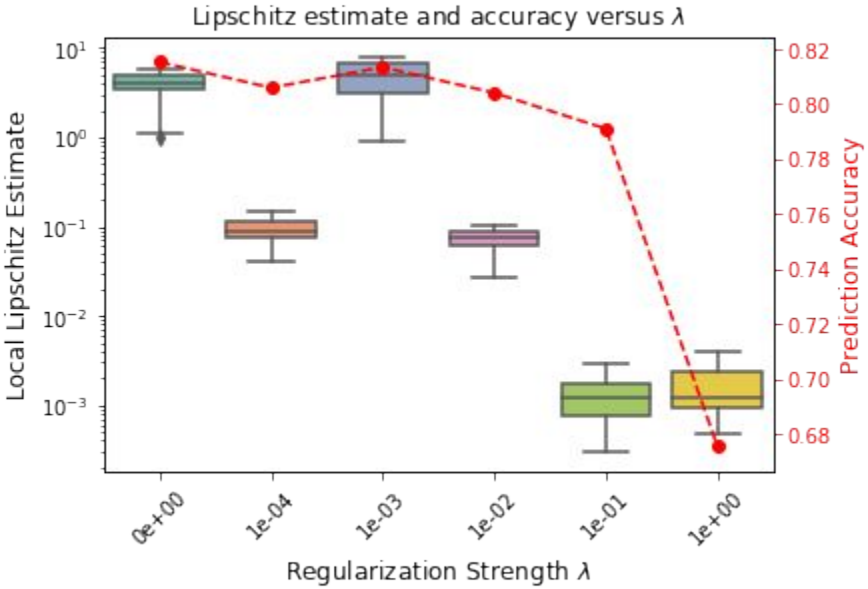
Input Value

Two_yr_Recidivism	0.0
Number_of_Priors	0.23
Age_Above_FourtyFive	0.0
Age_Below_TwentyFive	1.0
African_American	0.0
Asian	0.0
Hispanic	0.0
Native_American	0.0
Other	1.0
Female	0.0
Misdemeanor	0.0





Results - Stability



Conclusion and discussion

- The stability can be reproduced
- Accuracies mostly reproducible
- The explainability not in line with paper (When the conceptizer is used)
  - Not human interpretable

THANK YOU!

Questions?

Results - Faithfulness -Appendix