Explaining Latent Representations with a Corpus of Examples

LUH|AI

Leibniz Universität Hannover

Jonathan Crabbé, Zhaozhi Qian, Fergus Imrie, Mihaela van der Schaar

Poster Presentations in context of "Interpretable Machine Learning"

by: Jasmin Denk, Lukas Zain, Meike Liedtke

1 TL;DR

SimplEx:

- Post-hoc example-based explanations
- Creating a decomposition using latent representations and evaluating model reliability using corpus examples
- Extraction of positive or negative features influencing the models predictions ("jacobian projections")

2 Motivation & Problem Setting

Motivation

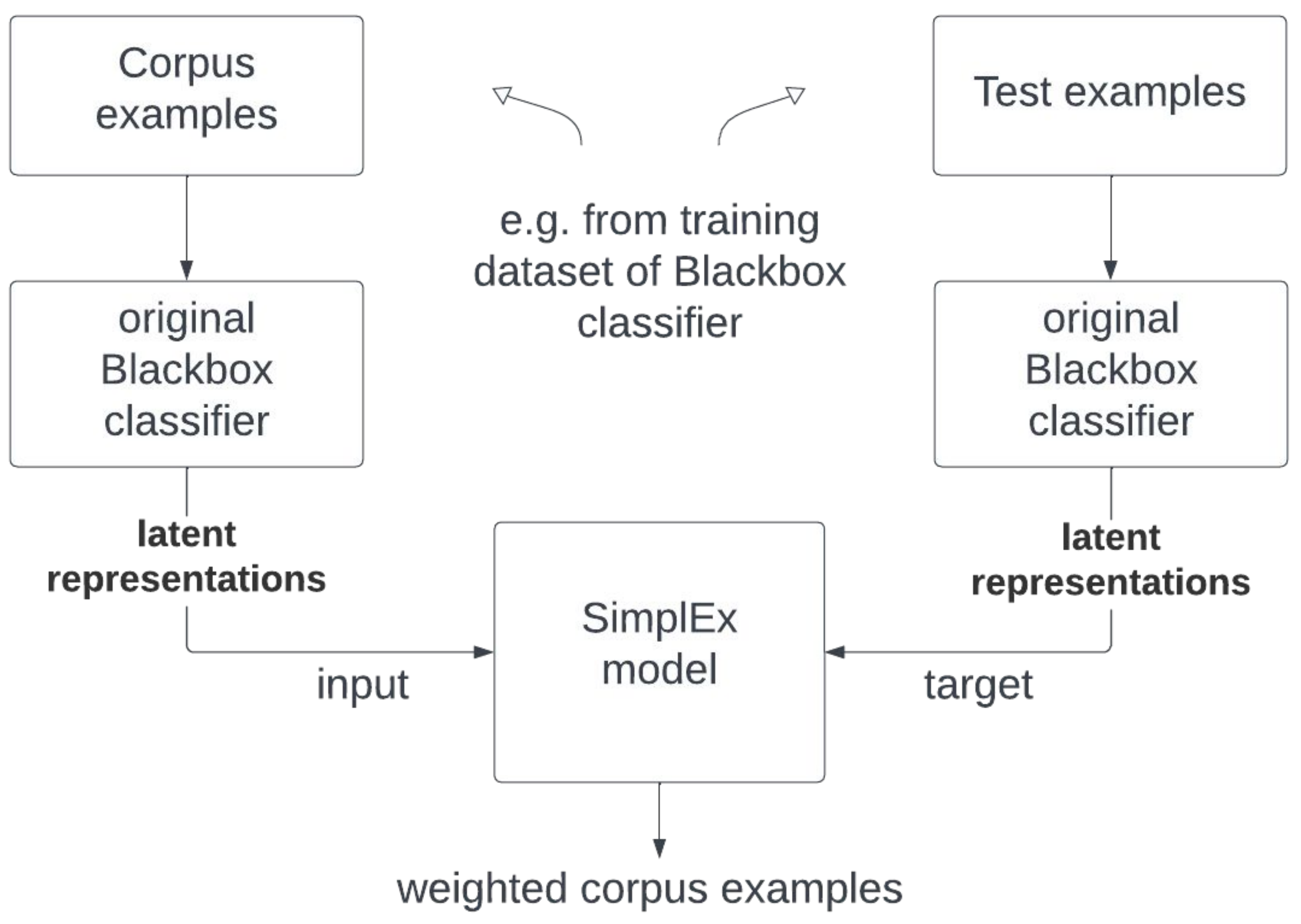
- Prediction credibility of black box classifiers is hard to interpret
- We want to gain insights by explaining them with a corpus of chosen examples

Problem Setting

How can we evaluate the credibility of a complex machine learning model, when using models interpretable by design is not available?

3 Approach

- Simplex trains a model using latent representations of corpus and test data
- Latent representation are vectors of the internal layers of a neural network
- Used latent representations are derived here from the last Layer of the original classifier
- Interpreting weights of corpus examples as percentual importance
- Jacobian projections: applying generalized integrated gradients to corpus examples to highlight feature contribution

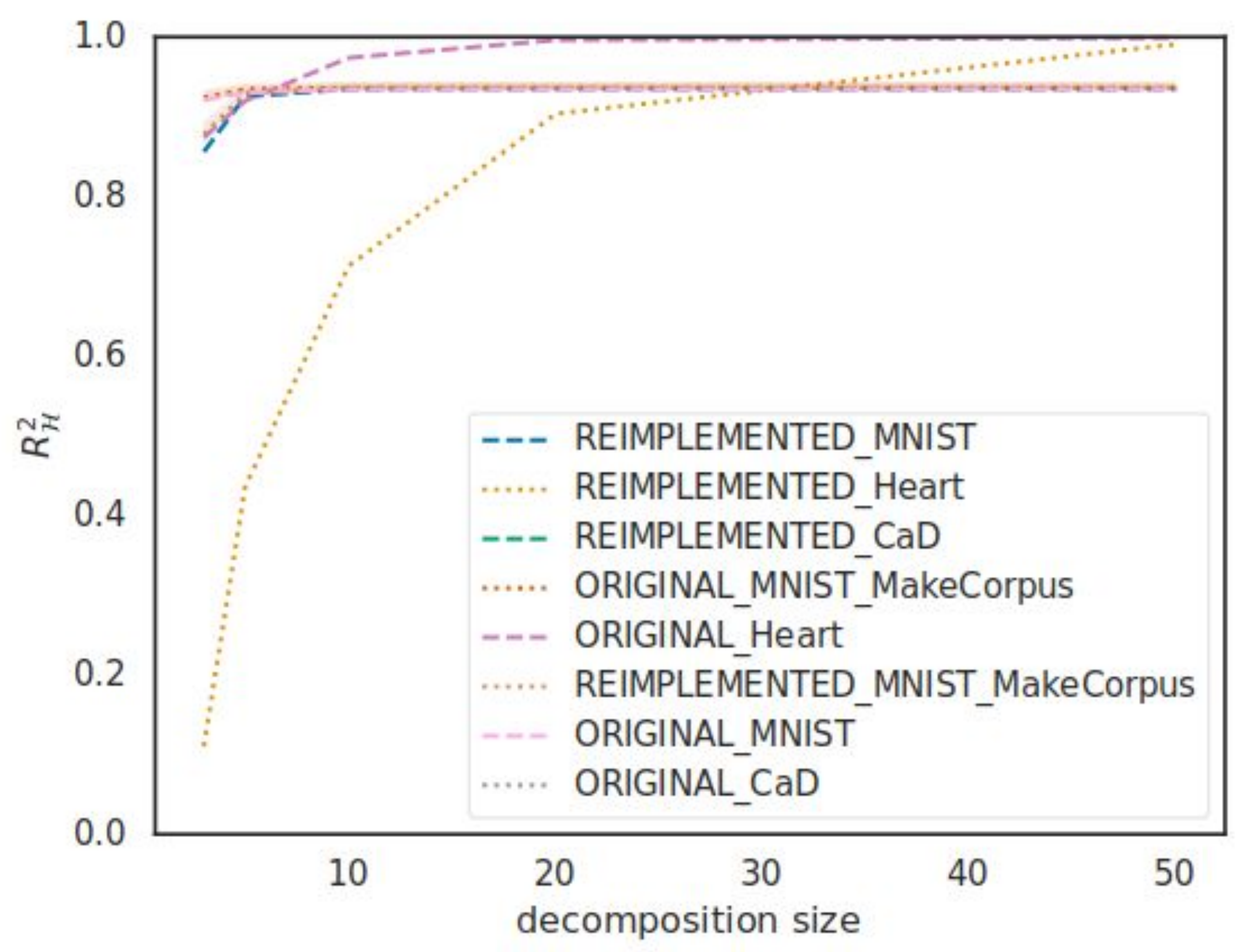


5 Future Works

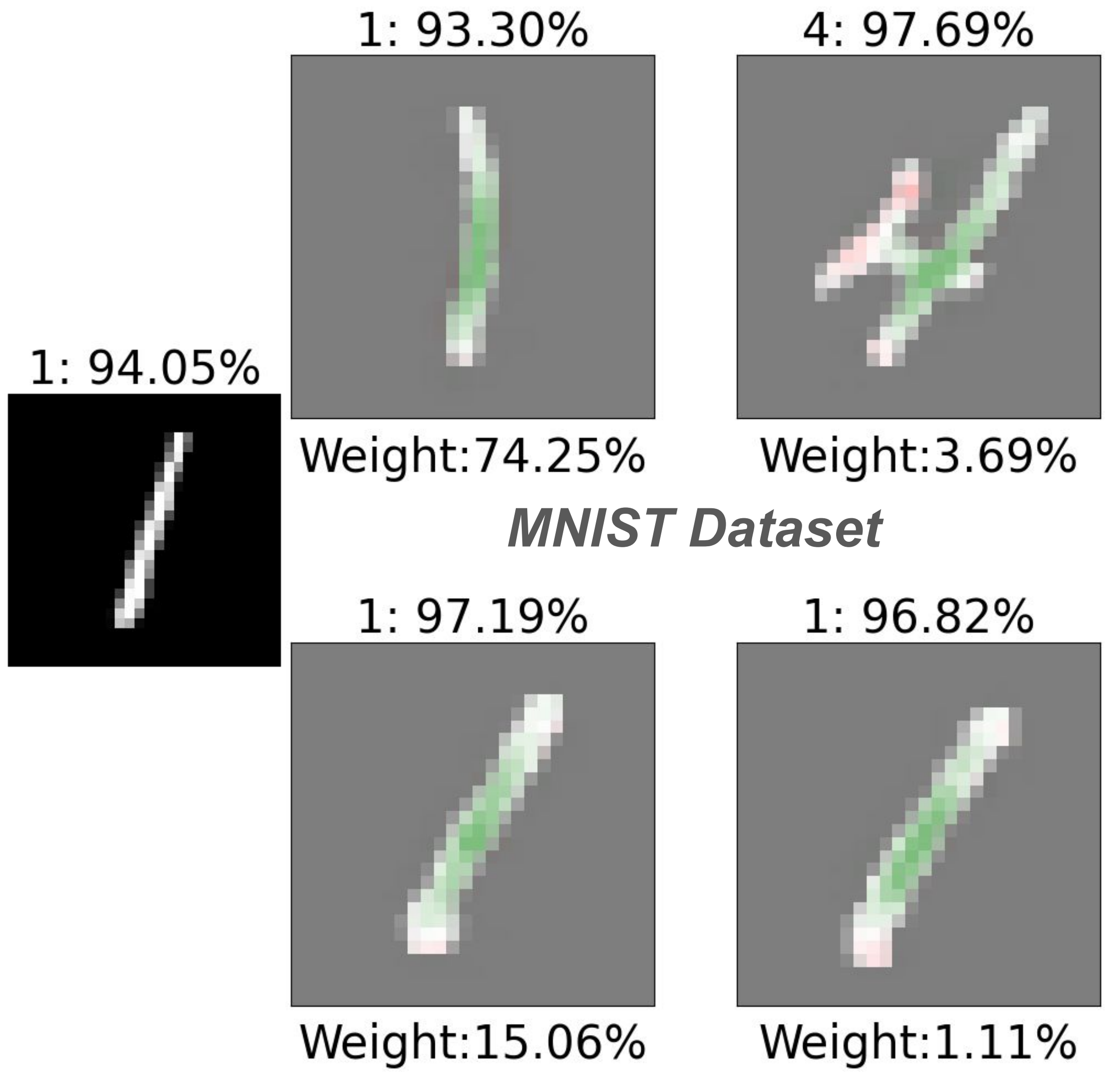
- Evaluating the role of the chosen corpus examples
- Examining security questions: can the original training dataset of the Blackbox Model be inferred?

4 Key Insights

- Re-implemented Model works almost as good as original simplex model
- Softmax Layer (or other normalization layer) is crucial during training

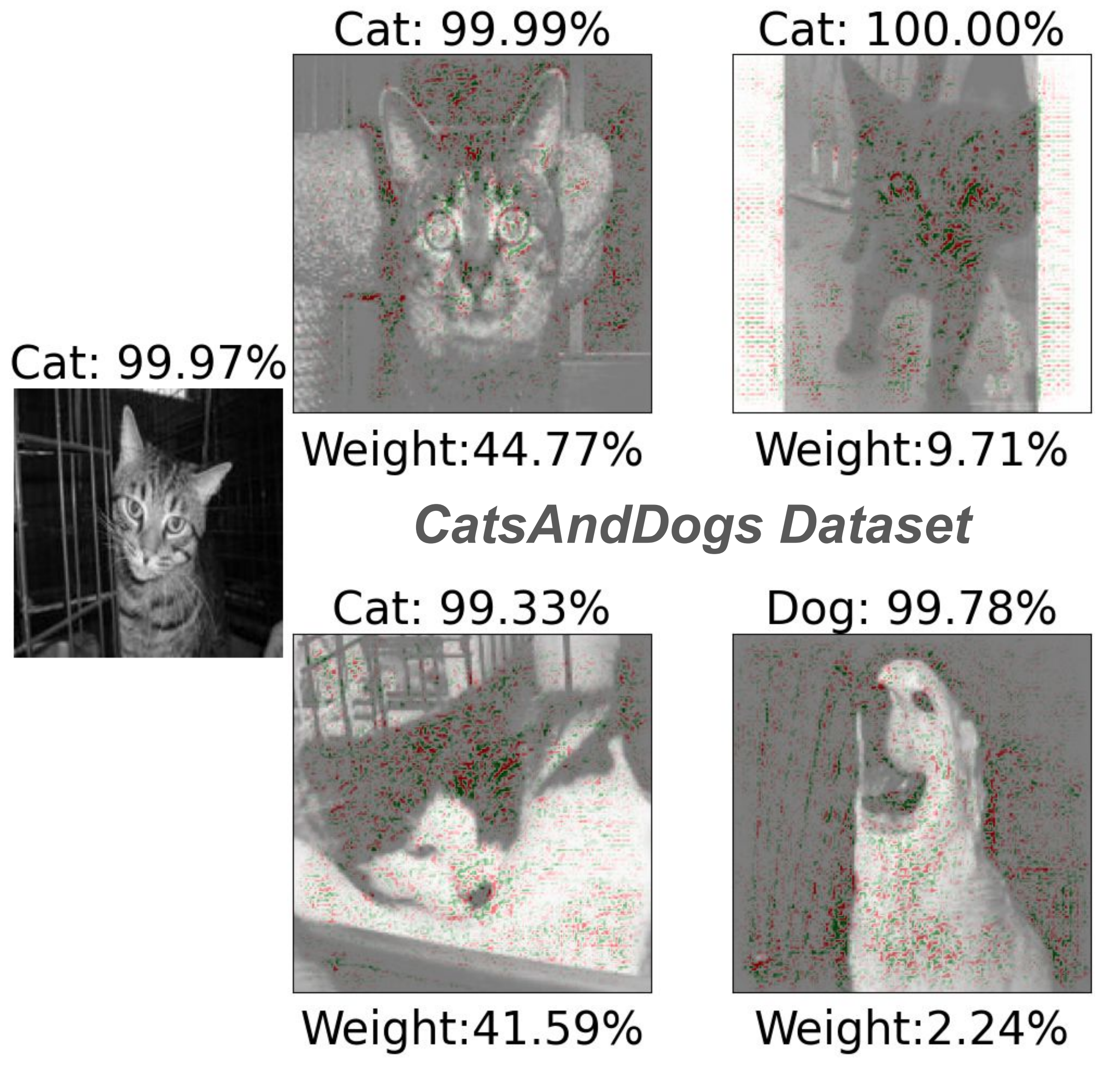


- Extraction of latent shapes can be difficult when using another classifier
- More complex input images produce less interpretable pixel-wise input attribution
- Positive and negative input attribution can be close
- R2 scores worse for more complex classifiers



MNIST Dataset

- Plausibility: When original test example is contained in the corpus, this sample makes up 99% of the explanation
- Corpus decomposition can include images from different classes
- New score for credibility: most important corpus examples must have same class as explained sample



CatsAndDogs Dataset