CXPlain: Causal Explanations for Model Interpretation under Uncertainty

Patrick Schwab (@schwabpa), and **Walter Karlen** (@mhsl_ethz)
Institute of Robotics and Intelligent Systems, ETH Zurich, Switzerland

1 Introduction

Feature importance estimates that inform users about the degree to which given inputs influence the output of a predictive model are crucial for **understanding**, **validating**, **and interpreting** machine-learning models.

How do we provide
- **accurate** importance scores **quickly**
- **for any model,** and
- estimate their **uncertainty?**

2 Causal explanations (CXPlain)

The main idea behind CXPlain is to train an **explanation model** to explain a given model (Figure 1). This framework has the advantage that we **do not need to retrain or adapt** the original model to explain its decisions.

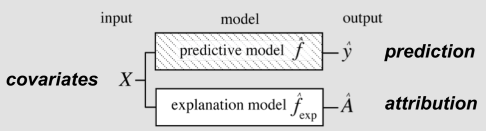


Figure 1. A conceptual overview of causal explanation models.

To train CXPlain, we transform the task of producing feature importance estimates for a given model into a **supervised learning task** by using a **causal objective** [1, 2].

$$\Delta\varepsilon_{X,i} = \varepsilon_{X\setminus\{i\}} - \varepsilon_X$$

3 Usage

1 TRAIN MODEL

```
x_train, y_train = ...
x_test = ...
model = ...

model.fit(x_train,
          y_train)
```

2 TRAIN CXPLAIN

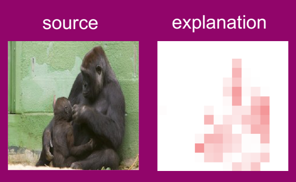
```
cxplain = CXPlain(
            model,
            builder,
            masking,
            loss
          )

cxplain.fit(x_train,
            y_train)
```

3 EXPLAIN MODEL

```
cxplain.explain(
            x_test
          )
```

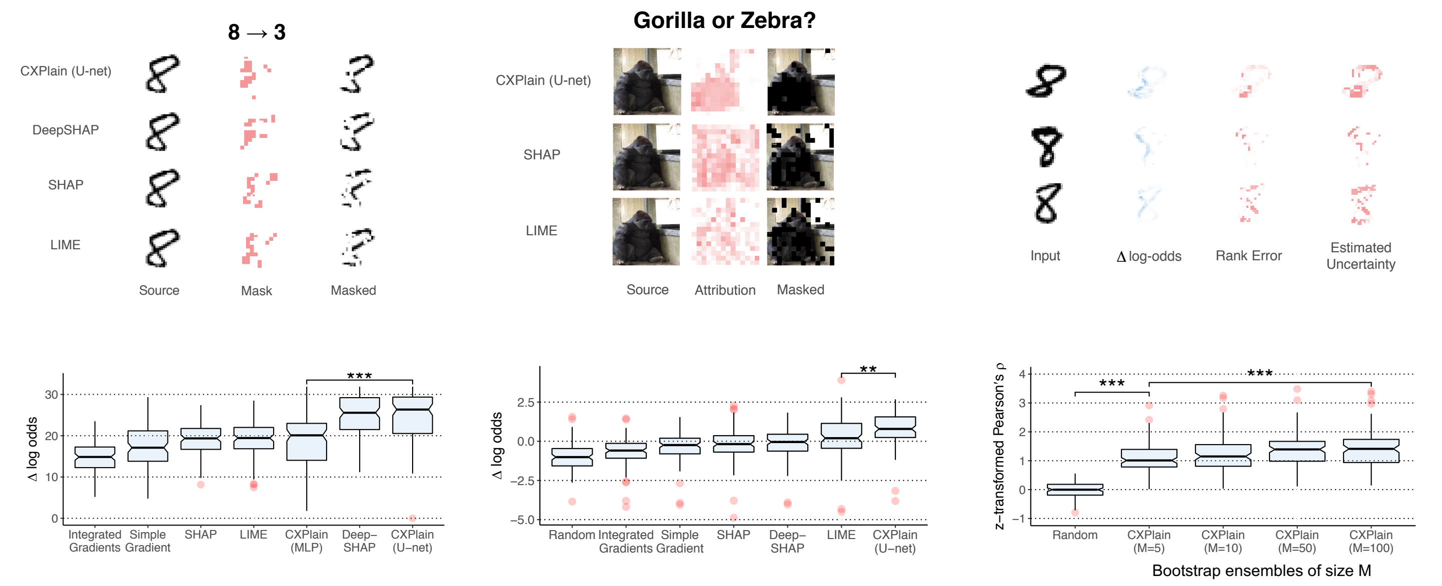
4 VISUALISE RESULT



TRY IT YOURSELF AT
github.com/d909b/cxplain

4 Results

MNIST [1,3] | **ImageNet** | **Uncertainty**



5 Components

Model Structure. In this work, we focus on. *neural explanation models*. However, in principle, any supervised model could be used.

Causal objective. We use a causal objective that quantifies the *marginal contribution of a feature towards the model's accuracy* [1, 2].

Masking Operation. We use a masking operation, such as zero masking [2,6], to *estimate each feature's marginal contribution.*

6 Conclusion

We presented CXPlain, a new **method for learning to estimate feature importance for any machine-learning model**.
We demonstrated that CXPlain is **fast at explanation time, accurate**, and that we are able to estimate its **attribution uncertainty** using bootstrap resampling.

7 References

1. Patrick Schwab, Djordje Miladinovic, and Walter Karlen. Granger-causal Attentive Mixtures of Experts: Learning Important Features with Neural Networks. In AAAI Conference on Artificial Intelligence, 2019
2. Erik Štrumbelj, Igor Kononenko, and M Robnik Šikonja. Explaining instance classifications with interactions of subsets of feature values. Data & Knowledge Engineering, 68(10):886–904, 2009.
3. Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Learning important features through propagating activation differences. International Conference of Machine Learning, 2017.
4. Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems, pages 4768–4777, 2017.
5. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.
6. Luisa M Zintgraf, Taco S Cohen, Tameem Adel, and Max Welling. Visualizing deep neural network decisions: Prediction difference analysis. In International Conference on Learning Representations, 2017.

NeurIPS | 2019

FNSNF Swiss National Science Foundation
75 NRP Big Data National Research Programme