A Causality Inspired Framework for Model Interpretation (Wu et al., 2023)

Presented by Shanmugapriya Kanagasabapathi

Problem Formulation

Predicting explanation masks which indicate the explanatory importance of input tokens for machine learning models

Contributions

- Definition of a unified causal lens for understanding existing model interpretation methods
- Introduction of CIMI to improve the causal insufficiency drawback of existing methods

Unified Causal Lens

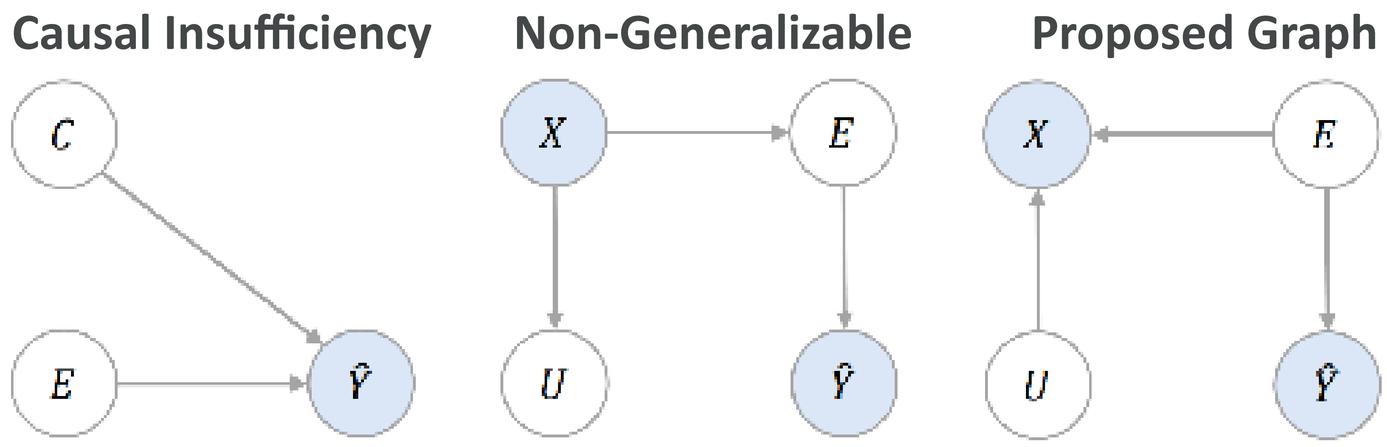

Figure 1: Comparison of causal graphs for model interpretation

Proposed Architecture for Model Interpretation

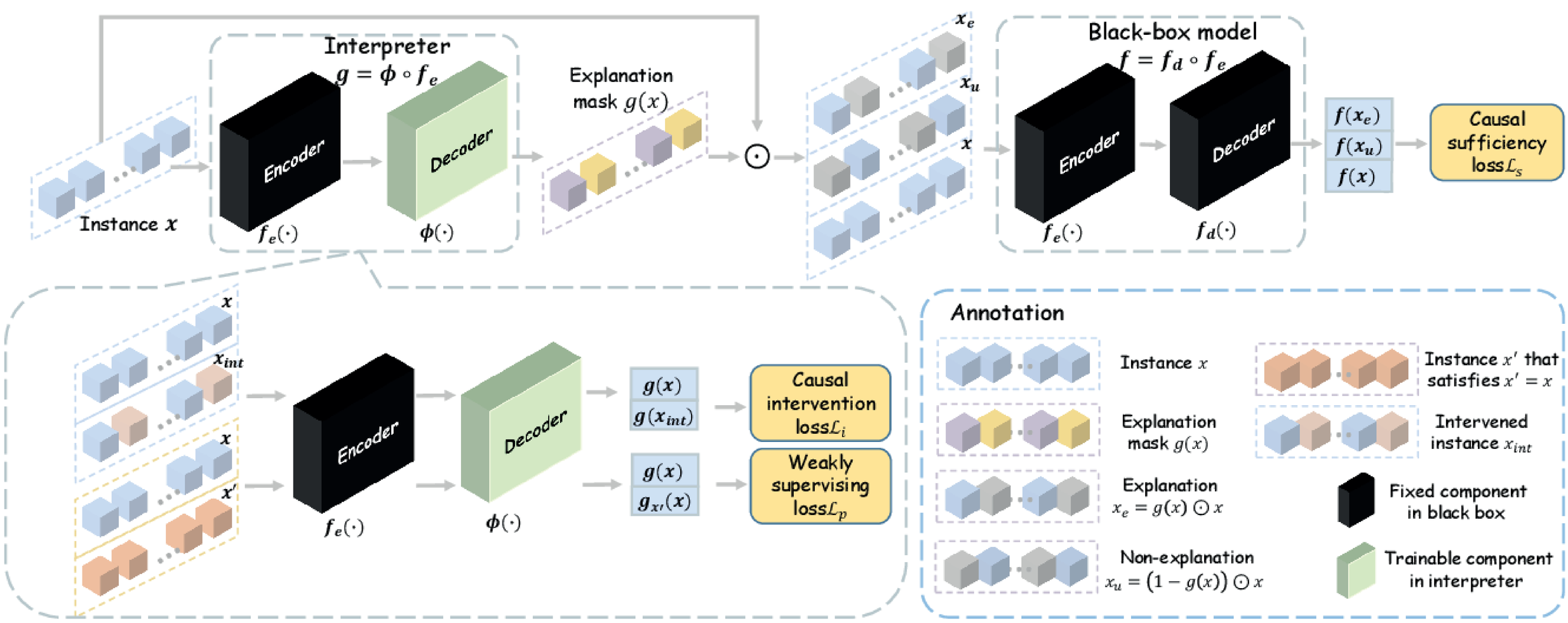

Figure 2: The framework of Causality Inspired Model Interpreter

Causal Sufficiency Loss $\mathcal{L}_s = \ell\left(f(x), f\left(x_e\right)\right) - \ell\left(f(x), f\left(x_u\right)\right)$

Causal Intervention Loss $\mathcal{L}_i = \ell\left(g(x), g\left(x_{int}\right)\right)$

Weakly Supervising Loss $\mathcal{L}_p = \log \sigma\left(g(x) - g_{x'}(x)\right)$

Overall Loss $\min_{\phi} \mathcal{L}_s + \alpha \mathcal{L}_i + \mathcal{L}_p$

Experimental Results: Comparison of Faithfulness | Generalizability | Effectiveness

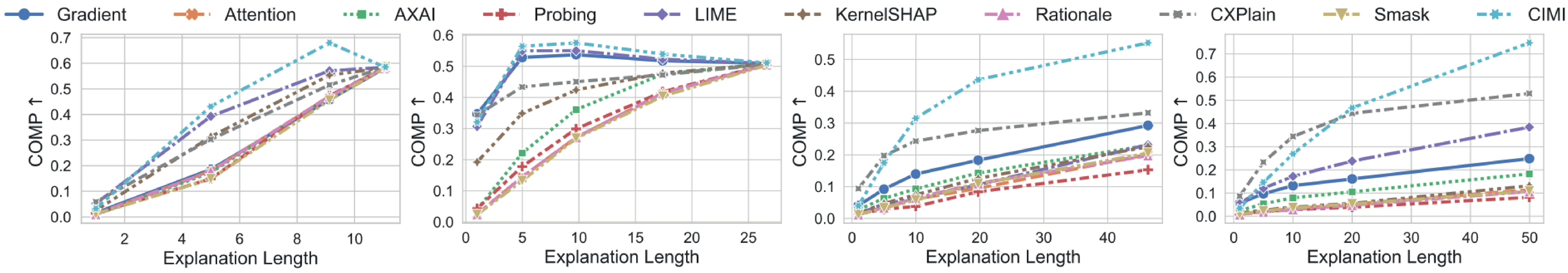

Figure 3: Comprehensiveness of different interpretability methods for different explanation lengths on Clickbait, Hate, Yelp and IMDB

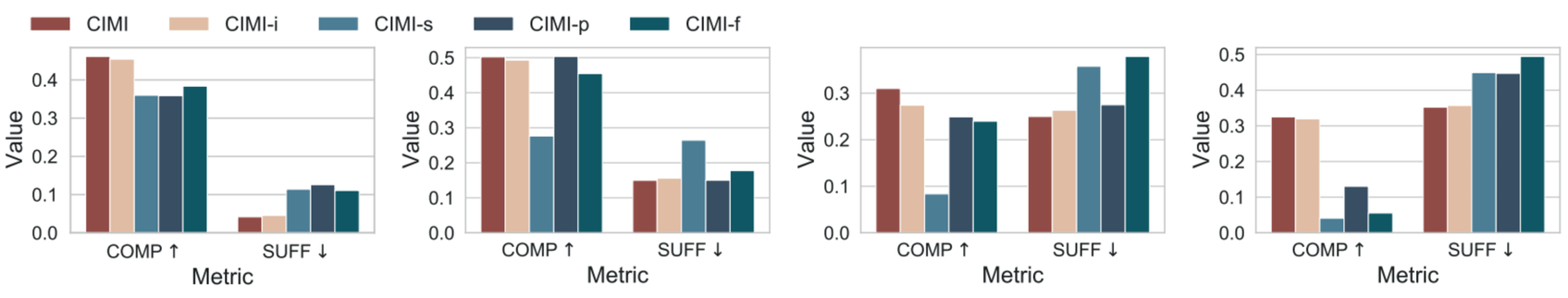

Figure 4: The faithful effect of the causal modules concerning comprehensivness and sufficiency on Clickbait, Hate, Yelp and IMDB

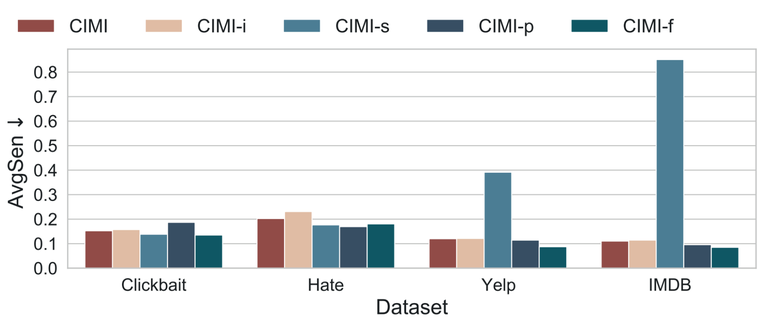

Figure 5: The generalizable effect of the causal modules concerning average sensitivity