10-23-2024

Developing Reliable Gradient Explanations for Artificial Intelligence: Addressing Consistency in Local Interpretability

Nolan M. Talaei
University of Massachusetts Lowell, nolan_talaei@uml.edu

Asil Oztekin
University of Massachusetts Lowell, asil_oztekin@uml.edu

Hongwei Zhu
University of Massachusetts Lowell, hongwei_zhu@uml.edu

Luvai Motiwalla
University of Massachusetts Lowell, luvai_motiwalla@uml.edu

Developing Reliable Gradient Explanations for Artificial Intelligence: Addressing Consistency in Local Interpretability

Research-in-Progress

Nolan M. Talaei
University of Massachusetts Lowell
nolan_talaei@uml.edu

Asil Oztekin
University of Massachusetts Lowell
asil_oztekin@uml.edu

Hongwei Zhu
University of Massachusetts Lowell
hongwei_zhu@uml.edu

Luvai Motiwalla
University of Massachusetts Lowell
luvai_motiwalla@uml.edu

ABSTRACT (REQUIRED)

Interpreting machine learning models remains challenging, particularly in high-stakes applications where trust and transparency are vital. We introduce Reliable Gradient Explanations (RGE), a method designed to enhance the stability and consistency of gradient-based feature importance explanations. RGE combines first-order gradient information with second-order Hessian elements to refine feature importance based on output curvature, reducing instability in traditional methods. Preliminary results indicate that RGE improves explanation accuracy and stability across different model architectures. Ongoing research aims to refine RGE, evaluate its performance on diverse datasets, and compare it with established interpretability techniques, ultimately promoting more transparent and reliable AI-driven decisions.

Keywords (Required)

Responsible Explainable AI (XAI), Stable Gradient-Based Interpretability, Feature Importance