

Integrated Gradients in Transformer Interpretation

Authors: Yongjin Cui, Xiaohui Fan, Xiaohui Fan

Date:

Abstract

Integrated Gradients (IG) addresses certain limitations of gradients in model interpretation and serves as an important tool in this field. This study analyzes and refutes three published Transformer interpretation methods based on IG. These incorrect methods have been published at top-tier academic conferences and, through their dissemination at these conferences, may have exerted adverse effects on both academia and industry. The publication of these erroneous methods can be attributed, on one hand, to an insufficient understanding of IG among researchers—including the paper authors and reviewers—and, on the other hand, to deficiencies in the research methodologies and evaluation practices within the model interpretation community. Our study carries no adversarial or targeted intent; its purpose is to eliminate the impact of these erroneous methods and to standardize the research methodologies and evaluation practices in this field. This paper focuses on three core aspects: why these methods are incorrect, why the erroneous methods were able to be published, and how to avoid similar mistakes in the future.

Note: Figure translations are in progress. See original paper for figures.

Source: ChinaXiv –Machine translation. Verify with original.