Interpretable Log Anomaly Detection

Benedek Huba Máth

Abstract

In the evolving landscape of large-scale computational systems, automated and interpretable log anomaly detection has become crucial for ensuring system reliability and security. This thesis addresses a key limitation of state-of-the-art log anomaly detection models: their lack of interpretability. While current models effectively identify anomalous sequences, they often fail to pinpoint the specific events responsible for the anomalies—information vital for system administrators and cybersecurity professionals. Building upon the PLELog model, this work proposes a novel interpretable framework that modifies the output layer to highlight anomalous log events within sequences. The approach integrates traditional log preprocessing (via the Drain parser), sequence modeling using GRUs, and an interpretable scoring mechanism based on λ statistics derived from TimeDistributed layers. The proposed model is evaluated on two benchmark datasets—HDFS and BGL—using both count matrix and embedding-based input representations. Results demonstrate that the new model maintains competitive performance while offering improved interpretability, with additional benefits observed in training efficiency and hyperparameter robustness. This work contributes a practical and explainable approach for industrial anomaly detection systems, with implications for both academic research and real-world deployments.