Efficient Anomaly Detection in Streaming Sensor Data Using Lightweight Transformer Models

Anonymous Author(s)
Affiliation Withheld for Review

Abstract

Real-time anomaly detection in streaming sensor data is critical for applications ranging from industrial monitoring to autonomous vehicles, yet most high-accuracy models are too computationally expensive for deployment on edge devices. In this work, we propose a lightweight transformer-based architecture optimized for low-latency anomaly detection in multivariate time series. Our model reduces parameter count by 68% compared to standard transformer baselines through a combination of sparse attention and depth-wise separable convolutions, while maintaining competitive detection accuracy. We evaluate our approach on three public benchmark datasets (SWaT, WADI, and SMD) and demonstrate that it achieves an F1 score within 2% of state-of-the-art methods while running 4.3x faster on embedded hardware. We further introduce a novel early-exit mechanism that allows the model to make confident predictions on clearly normal samples without executing the full network, reducing average inference time by an additional 31%. Our results suggest that carefully designed lightweight architectures can bring near-state-of-the-art anomaly detection capabilities to resource-constrained edge environments, enabling broader deployment of real-time monitoring systems in industrial and IoT settings.

1. Introduction

Anomaly detection in multivariate time series has become increasingly important as sensor networks proliferate across industrial control systems, smart infrastructure, and consumer IoT devices. While transformer-based architectures have demonstrated strong performance on this task, their computational cost typically restricts deployment to cloud or server-grade hardware, introducing latency that is unacceptable for many real-time safety-critical applications. This paper explores architectural modifications that preserve detection quality while substantially reducing inference cost, enabling deployment on constrained edge hardware such as embedded microcontrollers and low-power edge accelerators.

2. Method

Our architecture combines sparse local attention windows with depth-wise separable convolutional feature extractors, reducing both the parameter count and the computational complexity of standard self-attention. We additionally introduce an early-exit classification head that evaluates confidence at an intermediate layer, allowing the model to terminate computation early for inputs that are confidently classified as normal, while routing more ambiguous inputs through the full network depth.

3. Experiments

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We evaluate our method on three widely used benchmark datasets: SWaT (Secure Water Treatment), WADI (Water Distribution), and SMD (Server Machine Dataset). Across all three datasets, our lightweight model achieves F1 scores within 2% of larger transformer baselines, while reducing inference latency by 4.3x on representative embedded hardware. The early-exit mechanism provides an additional 31% reduction in average inference time with negligible impact on detection accuracy.

4. Conclusion

We presented a lightweight transformer architecture for real-time anomaly detection in streaming sensor data, demonstrating that careful architectural design can preserve near-state-of-the-art accuracy while meeting the latency and resource constraints of edge deployment. Future work will explore extending this approach to multi-modal sensor fusion and further compression via quantization.