REINVENTING PREDICTIVE MAINTENANCE:
Ultra-Optimized SRU Models for Embedded AI

Predictive maintenance, essential for anticipating failures and reducing downtime, often relies on complex models such as LSTMs (Long Short-Term Memory).
However, their high computational cost and resource requirements make them unsuitable for embedded systems.
Embedded architectures, such as VisionFive2 RISC-V, impose severe constraints in terms of latency, memory and energy, limiting their ability to integrate these models.
This article explores the optimization of LSTMs, their replacement by SRUs (Simple Recurrent Units), and the creation of a minimal version, to meet these challenges.

METHODOLOGY

Progressive model optimization

- Transformation of an LSTM into an SRU to reduce latency and power consumption.

- Use of techniques such as pruning (reduction of unnecessary connections) and knowledge distillation (compression with knowledge transfer).

- SRU Minimal: Structural simplification of SRUs by eliminating complex calculations that are unnecessary for long-term trends.
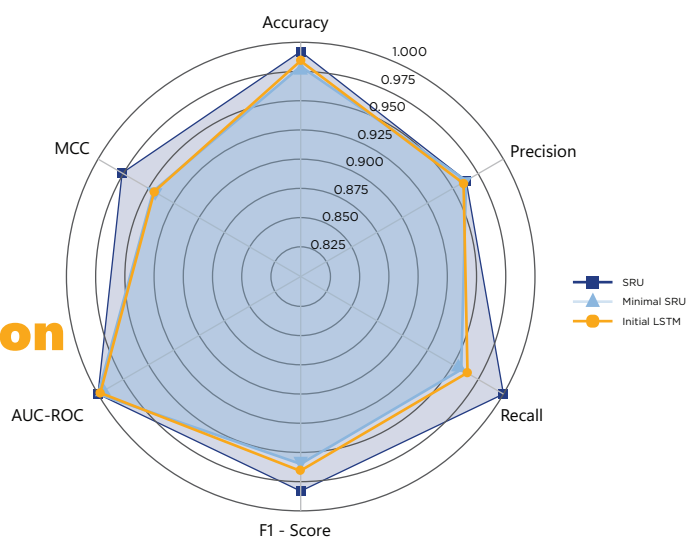

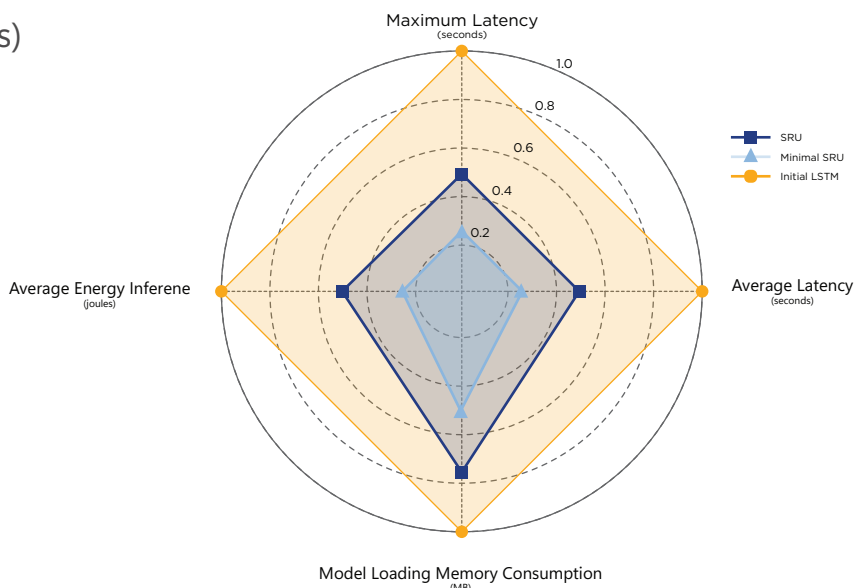Fig. 6: Predictive metric of initial LSTM and SRU models

Fig. 7: Enhancement of embedded performance of SRU models compared to the initial LSTM

Experimental evaluation

- Tests on a NASA dataset simulating the degradation of turbofan engines.
- Performance measurements (precision, recall, latency, energy).

MAIN RESULTS

- Reduced latency: SRU Minimal reduces latency by 75% while maintaining the same performance as the original LSTM.
- Energy optimization: up to 87% energy savings by inference
- Accuracy maintained or improved: Distilled SRU maintains an accuracy of 97.85%.

KEY APPLICATIONS FOR INDUSTRY

- IoT and autonomous systems: Reducing cloud requirements for embedded solutions.
- Aeronautics and heavy industry: maintenance of critical engines.

OUTLOOK

SRUs and Minimal SRUs mark a significant step forward for embedded systems. Future work will focus on designing customized SRU architectures to further improve accuracy and efficiency.
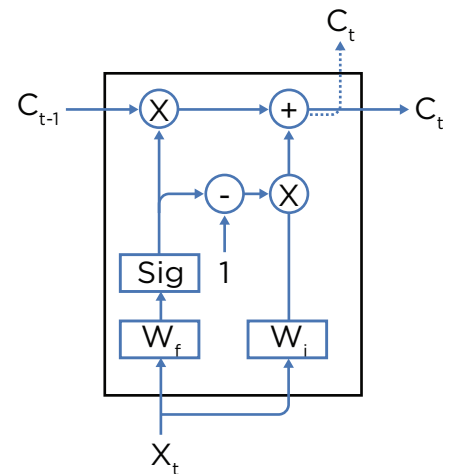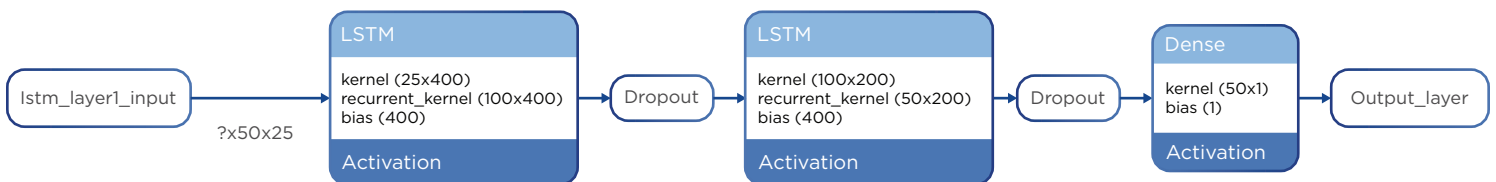
Fig. 3: Diagram of minimal SRU

Fig. 4: Model layers

Reference

(PDF) Optimizing SRU Models for Predictive Maintenance on Embedded Systems (RISC-V)