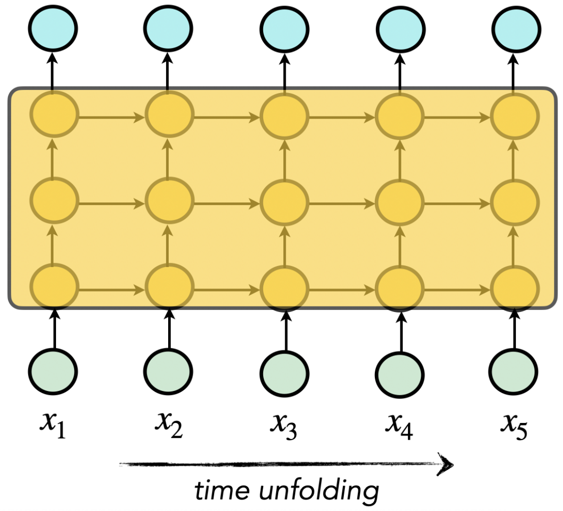
Quantized Recurrent Neural Networks for Efficient Long Sequence Modeling

Recurrent neural networks and state-space models are widely used for modeling long sequential data due to their ability to capture temporal dependencies with compact representations. However, their deployment in resource-constrained environments is often limited by the heavy computational cost and bandwidth requirements at inference time. Quantization offers a promising approach to reduce these requirements by representing model parameters and activations with lower precision. In this project, we will investigate quantized architectures for efficient long sequence modeling, focusing on the trade-offs between accuracy, stability, and hardware efficiency. In particular, we will explore how different quantization strategies and architectural choices impact performance and enable scalable inference.



Goals & Tasks

- Review of the state-of-the-art on quantization methods for neural networks.
- Implementation and benchmarking of quantized recurrent architectures for long sequence modeling.
- Evaluation of accuracy-efficiency trade-offs under different quantization schemes.

Contact

Ozan Özdenizci: oezdenizci@tugraz.at

Qualifications

- Interest in resource-efficient deep learning.
- Experience with the Python based deep learning framework PyTorch.
- Registered to one of the following:

☐ **Bachelor Thesis**
✓ **Seminar Project**
☐ **Master Thesis**