TECHNISCHE UNIVERSITÄT DARMSTADT

Thesis (B.Sc. / M.Sc.)
TL;DR (Too Long; Did Not Run) Factorized Temporal Attention for Processing Ultra Long Context

Department 18
Electrical Engineering and Information Technology
Self-Organizing Systems Lab

Prof. Dr. Heinz Koeppl
Head of lab

Philipp Froehlich
Project supervisor

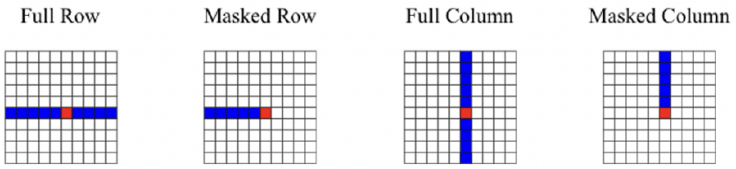
S3|06 206
Merckstrasse 25
64283 Darmstadt

philipp.froehlich@tu-darmstadt.de
https://www.bcs.tu-darmstadt.de

December 25, 2025

Background: Transformers have become a strong default for time series forecasting, but their quadratic attention cost makes *long temporal context* expensive in practice. Recent work in our group addressed this bottleneck by developing a factorized temporal attention mechanism that combines FlashAttention and *axial attention* for time series data. The factorization is motivated by the multi-scale periodic structure common in temporal signals: time can often be decomposed into interpretable axes such as *within-day vs. across-days*, *within-year vs. across-years*, or for human activity data *within-week patterns* (weekday/weekend effects) across multiple weeks. By attending along these axes separately, the model can capture long-range dependencies efficiently, enabling substantially longer histories (up to a 4× increase) without exceeding GPU memory limits and demonstrating stable training on irregular spatio-temporal data.

However, a key insight emerged when applying the same model to ERA5 weather forecasting: extending the context from 2 to 175 days did not yield physically meaningful accuracy gains (despite smoother and more stable predictions). This raises an important and practical research question: *When does long-context attention actually help?* The answer is crucial for guiding model development and for practitioners deciding whether long-context Transformers are worth the compute.



Visualization of Axial Attention [1]. Information can only be propagated within the respective axes.

Objective: The goal of this thesis is to determine *which forecasting problems benefit from extended temporal context* and *why*. Rather than applying the architecture indiscriminately, you will develop principled criteria for domain selection, run controlled experiments across multiple datasets, and analyze results to extract generalizable insights.

Key tasks (scope adaptable to interests):

- **Define long-context suitability criteria:** Identify domain characteristics that plausibly require long memory (e.g., delayed effects, multi-seasonal structure, regime shifts, rare/extreme events, slow dynamics, degradation). Translate these into testable hypotheses and measurable dataset descriptors.
- **Select and prepare 3–5 representative domains:** Build a diverse benchmark suite with strong justification for long-context needs. Candidate domains include energy demand and renewables (seasonal + weather-driven dependencies), epidemiology (delays and transmission dynamics), hydrology (catchment

memory), finance (regime changes and volatility clustering), climate indices (multi-year oscillations), and industrial condition monitoring (degradation and maintenance cycles).

- **Establish a rigorous comparison framework:** Evaluate factorized temporal attention at multiple context lengths against competitive baselines (standard Transformers, patching/downsampling approaches, classical forecasting methods where appropriate). Use metrics that reflect practical or physical relevance (e.g., event-based scores for extremes, domain-specific thresholds, calibration/stability measures), not only average error.

- **Analyze failure modes and scaling effects:** Investigate whether missing gains are due to insufficient model capacity, noise ceilings, weak long-range signal, or optimization issues. Where relevant, study scaling along depth/width and assess whether increased capacity unlocks benefits from longer contexts.

- **Deliver guidelines and recommendations:** Synthesize results into a set of heuristics and diagnostic tests that help practitioners decide whether long-context attention is likely to pay off for a given dataset and what context length / capacity trade-offs are sensible.

Prerequisites:

- Strong background in machine learning and deep learning or mathematical modelling.
- Proficiency in Python and PyTorch.
- Basic experiences with time series forecasting and experimental evaluation.
- Interest in careful scientific analysis: ablations, robustness checks, and domain-aware interpretation of results.

[1] Ho, J., Kalchbrenner, N., Weissenborn, D., & Salimans, T. (2019). Axial attention in multidimensional transformers. arXiv preprint arXiv:1912.12180.

For further information, please contact Philipp Froehlich.