INNOVATION PARTNERSHIPS
UNIVERSITY OF MICHIGAN

MonarchAttention: Zero-Shot Conversion to Fast, Hardware-Aware Structured Attention

TECHNOLOGY NUMBER: 2025-631

OVERVIEW

MonarchAttention is a novel software solution that makes deep learning models dramatically more efficient at processing long sequences, unlocking faster AI applications with minimal tradeoffs in accuracy.

- Instantly converts existing transformer models to a fast, hardware-optimized attention mechanism—no retraining required.
- Delivers up to 8x speedups for large inputs by efficiently approximating the attention matrix, reducing both compute and memory usage.

BACKGROUND

Transformers, the backbone of today’s top AI models in language, vision, and science, use an “attention” mechanism to understand how different parts of data relate to each other. This attention step powers breakthroughs in contextual understanding but is notoriously resource-intensive: as the input size (such as document or image length) grows, the time and memory required balloon quadratically. This limits the use of powerful AI on long texts, large images, or scientific data and makes deployment costly—despite a booming market demand for AI solutions that handle ever-longer context and data. Previous attempts to speed things up either sacrifice too much model fidelity, require costly retraining, or fail to deliver practical speed gains on real-world hardware.

Technology ID

2025-631

Category

Software

Software & Content

Inventor

Laura Balzano

Further information

Ashwathi Iyer

ashwathi@umich.edu

View online

INNOVATION

MonarchAttention sidesteps these pitfalls with a fundamentally new approach: it replaces the dense, bottleneck attention computation with a structured “Monarch matrix” optimized to mimic the full attention operation. Unlike traditional methods, MonarchAttention directly fits these matrices to the task at hand in a mathematically principled way, finding an efficient approximation—without any custom retraining. The software is specifically engineered to exploit modern GPUs, translating theoretical complexity savings into real speed: up to 1.4x-8x faster than the state of the art, depending on sequence length. Critically, the solution drops into existing transformer models, handling everything from text and code to images, and preserves model quality across diverse tasks. This establishes MonarchAttention as the first zero-shot, hardware-aware, accurate, and general solution to transformer acceleration.

ADDITIONAL INFORMATION

INTELLECTUAL PROPERTY:

Patent applications pending.