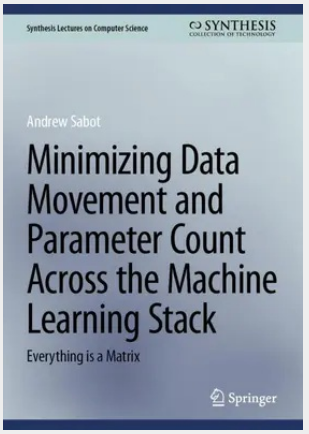
Sabot

Minimizing Data Movement and Parameter Count Across the Machine Learning Stack

Everything is a Matrix



This book provides a focused, research-forward guide to making large AI models efficient in practice and also presents an array of novel techniques to reduce memory footprint, accelerate computation, and improve overall hardware utilization. The author demonstrates that substantial efficiency gains can be achieved by rethinking how data is computed, stored, and compressed, with a special focus on matrices, the core computational structure underpinning both scientific computing and neural networks. Modern AI models run on huge grids of numbers (matrices/tensors), and their speed and affordability depend on how those numbers are arranged and processed on real hardware (GPUs/TPUs/CPUs). This book explains practical methods to skip unnecessary work (structured sparsity), move data efficiently (gather/scatter), and shrink models without losing accuracy (block distillation) so that AI systems can use less memory, less time, and less energy without sacrificing quality. In addition, the book shows how to turn algorithmic ideas into hardware-aware speedups on GPUs/TPUs. Readers will learn when sparsity pays off, how to schedule irregular workloads, and how to recover accuracy in compressed models. Case studies illustrate end-to-end design choices, evaluation, and pitfalls. The result is a coherent perspective that bridges theory, compilers/run times, and real-world deployment.

This book provides a focused, research-forward guide to making large AI models efficient in practice and also presents an array of novel techniques to reduce memory footprint, accelerate computation, and improve overall hardware utilization. The author demonstrates that substantial efficiency gains can be achieved by rethinking how data is computed, stored, and compressed, with a special focus on matrices, the core computational structure underpinning both scientific computing and neural networks. Modern AI models run on huge grids of numbers (matrices/tensors), and their speed and affordability depend on how those numbers are arranged and processed on real hardware (GPUs/TPUs/CPUs). This book explains practical methods to skip unnecessary work (structured sparsity), move data efficiently (gather/scatter), and shrink models without losing accuracy (block distillation) so that AI systems can use less memory, less time, and less energy without sacrificing quality. In addition, the book shows how to turn algorithmic ideas into hardware-aware speedups on GPUs/TPUs. Readers will learn when sparsity pays off, how to schedule irregular workloads, and how to recover accuracy in compressed models. Case studies illustrate end-to-end design choices, evaluation, and pitfalls. The result is a coherent perspective that bridges theory, compilers/run times, and real-world deployment.

42,79 €
39,99 € (zzgl. MwSt.)

vorbestellbar, Erscheinungstermin ca. Juni 2026

Artikelnummer: 9783032230997
Medium: Buch
ISBN: 978-3-032-23099-7
Verlag: Springer Nature Switzerland AG
Erscheinungstermin: 30.06.2026
Sprache(n): Englisch
Auflage: Erscheinungsjahr 2026
Serie: Synthesis Lectures on Computer Science
Produktform: Gebunden
Seiten: 110
Format (B x H): 168 x 240 mm

Kundenservice Verlag Dr. Otto Schmidt KG
Neumannstraße 10, 40235 Düsseldorf | kundenservice@otto-schmidt.de | 0221-93738-997 (Inland)