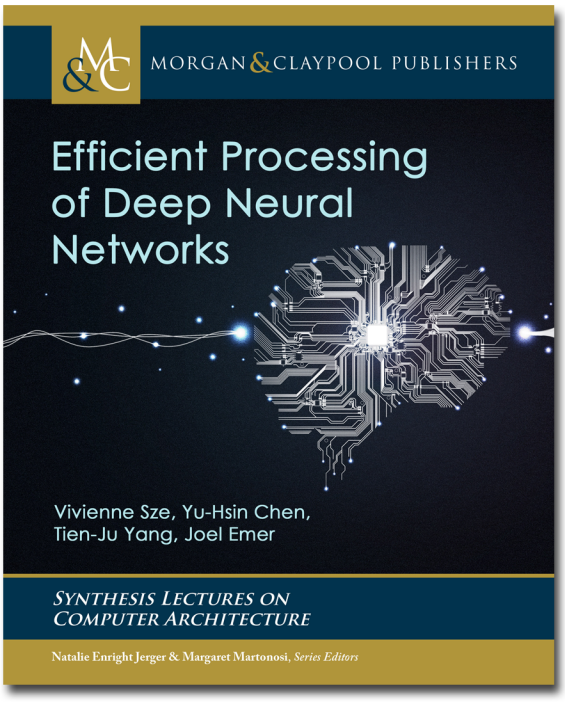
LIMITED TIME PRE-PUBLICATION PRICING

A structured treatment of the key principles and techniques for enabling efficient processing of deep neural networks (DNNs).



Efficient Processing of Deep Neural Networks

Vivienne Sze, *Massachusetts Institute of Technology*
Yu-Hsin Chen, *Massachusetts Institute of Technology*
Tien-Ju Yang, *Massachusetts Institute of Technology*
Joel S. Emer, *Massachusetts Institute of Technology, Nvidia Research*

Paperback ISBN: 9781681738543 • eBook ISBN: 9781681738550
Hardcover ISBN: 9781681738567 • May, 2020 • 75 pages
Paperback: $59.95 • eBook: $47.96 • Combo: $74.94
Hardcover $79.95 • Hardcover Combo $99.94

This book provides a structured treatment of the key principles and techniques for enabling efficient processing of deep neural networks (DNNs). DNNs are currently widely used for many artificial intelligence (AI) applications, including computer vision, speech recognition, and robotics. While DNNs deliver state-of-the-art accuracy on many AI tasks, it comes at the cost of high computational complexity. Therefore, techniques that enable efficient processing of deep neural networks to improve metrics—such as energy-efficiency, throughput, and latency—without sacrificing accuracy or increasing hardware costs are critical to enabling the wide deployment of DNNs in AI systems.

The book includes background on DNN processing; a description and taxonomy of hardware architectural approaches for designing DNN accelerators; key metrics for evaluating and comparing different designs; features of the DNN processing that are amenable to hardware/algorithm co-design to improve energy efficiency and throughput; and opportunities for applying new technologies. Readers will find a structured introduction to the field as well as a formalization and organization of key concepts from contemporary works that provides insights that may spark new ideas.

CONTENTS