1. Record Nr. UNINA9910831008703321

Autore	Zhang Baochang
Titolo	Neural Networks with Model Compression / / by Baochang Zhang, Tiancheng Wang, Sheng Xu, David Doermann
Pubbl/distr/stampa	Singapore : , : Springer Nature Singapore : , : Imprint : Springer, , 2024
ISBN	981-9950-68-6
Edizione	[1st ed. 2024.]
Descrizione fisica	1 online resource (267 pages)
Collana	Computational Intelligence Methods and Applications, , 2510-1773
Disciplina	006.32
Soggetti	Machine learning Artificial intelligence Image processing - Digital techniques Computer vision Machine Learning Artificial Intelligence Computer Imaging, Vision, Pattern Recognition and Graphics Computer Vision
Lingua di pubblicazione	Inglese
Formato	Materiale a stampa
Livello bibliografico	Monografia
Nota di bibliografia	Includes bibliographical references.
Nota di contenuto	Chapter 1. Introduction -- Chapter 2. Binary Neural Networks -- Chapter 3. Binary Neural Architecture Search -- Chapter 4. Quantization of Neural Networks -- Chapter 5. Network Pruning -- Chapter 6. Applications.
Sommario/riassunto	Deep learning has achieved impressive results in image classification, computer vision and natural language processing. To achieve better performance, deeper and wider networks have been designed, which increase the demand for computational resources. The number of floating-point operations (FLOPs) has increased dramatically with larger networks, and this has become an obstacle for convolutional neural networks (CNNs) being developed for mobile and embedded devices. In this context, our book will focus on CNN compression and acceleration, which are important for the research community. We will describe numerous methods, including parameter quantization, network pruning, low-rank decomposition and knowledge distillation. More recently, to reduce the burden of handcrafted architecture design,

neural architecture search (NAS) has been used to automatically build neural networks by searching over a vast architecture space. Our book will also introduce NAS due to its superiority and state-of-the-art performance in various applications, such as image classification and object detection. We also describe extensive applications of compressed deep models on image classification, speech recognition, object detection and tracking. These topics can help researchers better understand the usefulness and the potential of network compression on practical applications. Moreover, interested readers should have basic knowledge about machine learning and deep learning to better understand the methods described in this book.