1. Record Nr. UNINA9911020369203321

Autore	Raj Pethuru
Titolo	Model Optimization Methods for Efficient and Edge AI : Federated Learning Architectures, Frameworks and Applications
Pubbl/distr/stampa	Newark : , : John Wiley & Sons, Incorporated, , 2024 ©2025
ISBN	9781394219230 1394219237 9781394219223 1394219229 9781394219209 1394219202
Edizione	[1st ed.]
Descrizione fisica	1 online resource (428 pages)
Altri autori (Persone)	RahmaniAmir Masoud ColbyRobert NagasubramanianGayathri RanganathSunku
Disciplina	006.31
Soggetti	Edge computing Artificial intelligence
Lingua di pubblicazione	Inglese
Formato	Materiale a stampa
Livello bibliografico	Monografia
Nota di contenuto	Cover -- Title Page -- Copyright -- Contents -- About the Editors -- List of Contributors -- Chapter 1 Fundamentals of Edge AI and Federated Learning -- 1.1 Introduction -- 1.2 Concepts and Fundamentals of Edge AI -- 1.2.1 Defining Edge AI -- 1.2.2 Advantages of Edge AI -- 1.2.3 Challenges of Edge AI -- 1.2.4 Taxonomy of Edge AI Applications -- 1.3 Concepts and Fundamentals of FL -- 1.3.1 Defining FL -- 1.3.2 Advantages of FL -- 1.3.3 Challenges of FL -- 1.3.4 Taxonomy of FL Applications -- 1.4 Combining FL and Edge AI -- 1.5 Background -- 1.6 Applications of Edge AI and FL -- 1.7 Challenges, Future Research Directions, and Solutions -- 1.8 Conclusion -- References -- Chapter 2 AI Applications - Computer Vision and Natural Language Processing -- 2.1

Introduction -- 2.1.1 What Is It? Everything to Know About Artificial Intelligence -- 2.1.1.1 Super AI -- 2.1.1.2 General AI -- 2.1.1.3 Narrow AI -- 2.1.2 Key Ideas and Elements in the Study of Artificial Intelligence -- 2.1.3 Today's Modern Role of Artificial Intelligence -- 2.1.3.1 In the AI Basket, What Is There? -- 2.1.3.2 AI's Growth in India -- 2.1.3.3 Importance of AI in the Modern Era -- 2.1.3.4 AI's Effects on Business -- 2.1.3.5 How Can One Search for AI Jobs? -- 2.2 Artificial Intelligence: 2023's Top 18 Applications -- 2.2.1 AI Application in E Commerce -- 2.2.2 Artificial Intelligence Applications in Education -- 2.2.3 Artificial Intelligence Applications in Lifestyle -- 2.2.4 Navigational Applications of Artificial Intelligence -- 2.2.5 Applications of Artificial Intelligence in Robotics -- 2.2.6 Artificial Intelligence Applications in Human Resources -- 2.2.7 Artificial Intelligence Applications in Healthcare -- 2.2.8 Applications of Artificial Intelligence in Agriculture -- 2.2.9 Artificial Intelligence Applications in Gaming. 2.2.10 Artificial Intelligence Applications in Automobiles -- 2.2.11 Applications of Artificial Intelligence in Social Media -- 2.2.12 Applications of Artificial Intelligence in Marketing -- 2.2.13 Artificial Intelligence Applications for Chatbots -- 2.2.14 Artificial Intelligence Applications in Finance -- 2.2.15 AI in Astronomical Science -- 2.2.16 AI for Data Security -- 2.2.17 AI in Transportation and Travel -- 2.2.18 AI in the Automobile Industry -- 2.3 AI in Computer Visions -- 2.4 AI in Natural Language Processing -- 2.5 Conclusion -- 2.6 AI's Opportunities for Computer Vision and Natural Language Processing -- References -- Chapter 3 An Overview of AI Platforms, Frameworks, Libraries, and Processors -- 3.1 Introduction: Artificial Intelligence Platforms, Frameworks, Libraries, and Processors: The Building Blocks of AI Development -- 3.2 Edge AI -- 3.2.1 AI Platforms -- 3.2.1.1 AIOps -- 3.2.2 AI Frameworks -- 3.2.3 AI Libraries -- 3.2.3.1 MLIR in the AI Domain -- 3.2.4 AI Processors -- 3.2.4.1 Conclusion -- 3.2.4.2 Specialty ASIC Hardware for AI Training and Inference Workloads -- 3.2.4.3 Hardware Selection for AI Solutions for Cloud and Edge Use Cases -- References -- Chapter 4 Model Optimization Techniques for Edge Devices -- 4.1 Overview of Model Optimization Techniques -- 4.1.1 Predeployment Optimization Techniques -- 4.1.2 Deployment Time Optimization Techniques -- 4.1.3 Postdeployment Optimization Techniques -- 4.1.4 Examples -- 4.2 Deep Dive of Predeployment Model Optimization Techniques -- 4.2.1 Model Architecture Selection -- 4.2.1.1 Vision Model Architectures -- 4.2.1.2 Lightweight Transformers -- 4.2.1.3 Lightweight LLMs -- 4.2.2 Quantization -- 4.2.3 Structured Pruning -- 4.2.4 Knowledge Distillation -- 4.2.5 Sparsification -- 4.3 Deep Dive of DeploymentTime Model Optimization Techniques.
4.3.1 Conversion to Intermediate Representation -- 4.3.2 Graph Optimizations -- 4.3.3 TargetDependent Optimizations -- 4.3.4 Dynamic Batching -- 4.3.5 Model Caching -- 4.3.5.1 InMemory Model Caching -- 4.3.5.2 OnDisk Model Caching -- 4.3.6 Model Parallelism -- 4.4 Deep Dive of PostDeployment Model Optimization Techniques -- 4.4.1 Model Monitoring -- 4.4.2 Model Retraining -- 4.4.3 Hardware Upgrades -- 4.4.4 Feedback Loops -- 4.5 Summary -- References -- Chapter 5 AI Model Optimization Techniques -- 5.1 Introduction -- 5.1.1 Model Optimization -- 5.1.2 Essentiality of AI Model Optimization -- 5.1.3 Categories of Optimization Techniques -- 5.1.3.1 Computational Optimization -- 5.1.3.2 Efficiency Optimization -- 5.2 Pruning -- 5.2.1 Identifying and Eliminating Inessential Connections -- 5.2.2 Benefits of Pruning -- 5.2.2.1 Model Size Reduction -- 5.2.2.2 Accelerated Inference -- 5.2.3 Pruning Strategies -- 5.2.3.1 MagnitudeBased Pruning -- 5.2.3.2 Structured Pruning --

5.2.3.3 Iterative Pruning -- 5.2.4 Caveats and Considerations -- 5.2.5 RealWorld Applications -- 5.3 Quantization -- 5.3.1 Need for Quantization -- 5.3.2 PostTraining Quantization -- 5.3.2.1 Weight Quantization -- 5.3.2.2 Activation Quantization -- 5.3.2.3 QuantizationAware Training -- 5.3.3 Quantization Schemes -- 5.3.3.1 FixedPoint Quantization -- 5.3.3.2 Dynamic Range Quantization -- 5.3.3.3 Hybrid Quantization -- 5.3.4 TradeOff: Precision vs. Efficiency -- 5.3.5 RealWorld Applications -- 5.4 Model Distillation -- 5.4.1 Essence of Model Distillation -- 5.4.2 Knowledge Transfer -- 5.4.3 Benefits of Model Distillation -- 5.4.3.1 Compact Models -- 5.4.3.2 Efficient Inference -- 5.4.3.3 Improved Generalization -- 5.4.3.4 Adaptability -- 5.4.4 Practical Applications -- 5.5 Layer Fusion -- 5.5.1 Rationale Behind Layer Fusion -- 5.5.2 Depthwise Separable Convolutions.
5.5.3 Convolutional Layer Fusion -- 5.5.4 RealWorld Applications -- 5.6 Parallelization -- 5.6.1 Imperative of Parallelization -- 5.6.2 Data Parallelism -- 5.6.3 Model Parallelism -- 5.6.4 Pipeline Parallelism -- 5.6.5 RealWorld Applications -- 5.7 Hardware Acceleration -- 5.7.1 The Role of Hardware Acceleration -- 5.7.2 Optimizing for Specific Hardware -- 5.7.3 Impact of Hardware Acceleration -- 5.7.4 RealWorld Implementations -- 5.8 Transfer Learning -- 5.8.1 Essence of Transfer Learning -- 5.8.2 Benefits of Transfer Learning -- 5.8.2.1 Reduced Training Time -- 5.8.2.2 Enhanced Performance -- 5.8.2.3 Lower Data Requirements -- 5.8.2.4 Transfer of Knowledge -- 5.8.3 RealWorld Applications -- 5.8.4 The FineTuning Process -- 5.9 Neural Architecture Search -- 5.9.1 Essence of Neural Architecture Search -- 5.9.2 Benefits of Neural Architecture Search -- 5.9.2.1 TaskSpecific Architectures -- 5.9.2.2 Reduced Human Bias -- 5.9.2.3 Resource Efficiency -- 5.9.2.4 StateoftheArt Results -- 5.9.3 RealWorld Applications -- 5.9.4 NAS Techniques -- 5.9.5 Challenges in NAS -- 5.10 Pragmatic Optimization -- 5.10.1 Pragmatic Approach -- 5.10.2 TradeOffs in Pragmatic Optimization -- 5.10.3 RealWorld Applications -- 5.11 Conclusion -- References -- Chapter 6 Federated Learning: Introduction, Evolution, Working, Advantages, and Its Application in Various Domains -- 6.1 Introduction to Machine Learning and Federated Learning -- 6.2 Evolution of Federated Learning -- 6.3 How Federated Learning Works -- 6.4 Some Scholarly Work Related to Federated Learning -- 6.5 Distinct Advantages of Federated Learning -- 6.5.1 Federated Learning is Collaborative in Nature -- 6.5.2 Time Saving -- 6.5.3 Security -- 6.5.4 It Generates More Diverse Data -- 6.5.5 It Generates RealTime Prediction -- 6.5.6 It Is HandsOff and Noninvasive -- 6.5.7 Hardware Efficiency.
6.5.8 EdgeComputing Capabilities -- 6.5.9 Continuous Learning and Personalization -- 6.5.10 Scalability and Flexibility -- 6.6 Applications of Federated Machine Learning -- 6.6.1 Healthcare -- 6.6.2 Finance -- 6.6.3 Internet of Things -- 6.6.4 Autonomous Vehicles -- 6.6.5 Natural Language Processing -- 6.6.6 Federated Healthcare Models -- 6.6.7 Personalized Recommendations -- 6.6.8 Defence and National Security -- 6.6.9 Federated Learning in Education -- 6.6.10 Resource Allocation and Planning -- 6.6.11 Edge Computing -- 6.6.12 Research Collaborations -- 6.6.13 Personal Health Assistants -- 6.6.14 Federated Learning in Energy Sector -- 6.6.15 Supply Chain and Manufacturing -- 6.6.16 Environmental Monitoring -- 6.6.17 Federated Learning in Law Enforcement -- 6.6.18 Secure Communication -- 6.7 Conclusion -- References -- Chapter 7 Application Domains of Federated Learning -- 7.1 Introduction -- 7.1.1 Overview of Federated Learning -- 7.1.2 Importance and Advantages of Federated Learning -- 7.1.2.1 Why Federated Learning Is Crucial Right Now? -- 7.1.2.2

Advantages of Federated Learning -- 7.1.3 Objectives of the Chapter -- 7.2 Healthcare -- 7.2.1 Patient Privacy and Data Sharing Challenges -- 7.2.2 Federated Learning for Clinical Decision Support Systems -- 7.2.3 Federated Learning for Disease Diagnosis and Prediction -- 7.2.4 Challenges and Opportunities in Healthcare Data Federations -- 7.2.4.1 Challenges -- 7.2.4.2 Opportunities -- 7.3 Finance and Banking -- 7.3.1 Privacy and Security Concerns in Financial Data -- 7.3.2 Fraud Detection and Prevention with Federated Learning -- 7.3.3 Personalized Financial Services and Recommendations -- 7.3.4 Compliance and Risk Management in Federated Environments -- 7.3.4.1 How Federated Learning Functions for Compliance and Risk Management Is Explained as Follows.
7.3.4.2 Some Advantages of Federated Learning for Compliance and Risk Management Are Listed as Follows.

Sommario/riassunto

Comprehensive overview of the fledgling domain of federated learning (FL), explaining emerging FL methods, architectural approaches, enabling frameworks, and applications Model Optimization Methods for Efficient and Edge AI explores AI model engineering, evaluation, refinement, optimization, and deployment across multiple cloud environments (public, private, edge, and hybrid). It presents key applications of the AI paradigm, including computer vision (CV) and Natural Language Processing (NLP), explaining the nitty-gritty of federated learning (FL) and how the FL method is helping to fulfill AI model optimization needs. The book also describes tools that vendors have created, including FL frameworks and platforms such as PySyft, Tensor Flow Federated (TFF), FATE (Federated AI Technology Enabler), Tensor/IO, and more. The first part of the text covers popular AI and ML methods, platforms, and applications, describing leading AI frameworks and libraries in order to clearly articulate how these tools can help with visualizing and implementing highly flexible AI models quickly. The second part focuses on federated learning, discussing its basic concepts, applications, platforms, and its potential in edge systems (such as IoT). Other topics covered include: * Building AI models that are destined to solve several problems, with a focus on widely articulated classification, regression, association, clustering, and other prediction problems * Generating actionable insights through a variety of AI algorithms, platforms, parallel processing, and other enablers * Compressing AI models so that computational, memory, storage, and network requirements can be substantially reduced * Addressing crucial issues such as data confidentiality, data access rights, data protection, and access to heterogeneous data * Overcoming cyberattacks on mission-critical software systems by leveraging federated learning Written in an accessible manner and containing a helpful mix of both theoretical concepts and practical applications, Model Optimization Methods for Efficient and Edge AI is an essential reference on the subject for graduate and postgraduate students, researchers, IT professionals, and business leaders.