Model Compression and Hardware Acceleration for Neu
Survey

Proceedings of the IEEE

108, 485-532

DOI: 10.1109/jproc.2020.2976475

Citation Report

#	Article	IF	Citations
1	Hybrid tensor decomposition in neural network compression. Neural Networks, 2020, 132, 309-320.	3.3	25
2	An Updated Survey of Efficient Hardware Architectures for Accelerating Deep Convolutional Neural Networks. Future Internet, 2020, 12, 113.	2.4	111
3	Compressing 3DCNNs based on tensor train decomposition. Neural Networks, 2020, 131, 215-230.	3.3	18
4	A birdâ€™s-eye view of deep learning in bioimage analysis. Computational and Structural Biotechnology Journal, 2020, 18, 2312-2325.	1.9	94
5	HFNet: A CNN Architecture Co-designed for Neuromorphic Hardware With a Crossbar Array of Synapses. Frontiers in Neuroscience, 2020, 14, 907.	1.4	13
6	Recent Progress on Memristive Convolutional Neural Networks for Edge Intelligence. Advanced Intelligent Systems, 2020, 2, 2000114.	3.3	19
7	A Systematic Study of Tiny YOLO3 Inference: Toward Compact Brainware Processor With Less Memory and Logic Gate. IEEE Access, 2020, 8, 142931-142955.	2.6	15
8	Hardware and Software Optimizations for Accelerating Deep Neural Networks: Survey of Current Trends, Challenges, and the Road Ahead. IEEE Access, 2020, 8, 225134-225180.	2.6	91
9	Brain-Inspired Computing: Models and Architectures. IEEE Open Journal of Circuits and Systems, 2020, 1, 185-204.	1.4	21
10	Hardware Implementation of Deep Network Accelerators Towards Healthcare and Biomedical Applications. IEEE Transactions on Biomedical Circuits and Systems, 2020, 14, 1138-1159.	2.7	93
11	Learning Slimming SAR Ship Object Detector Through Network Pruning and Knowledge Distillation. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14, 1267-1282.	2.3	58
12	MOSDA: On-Chip Memory Optimized Sparse Deep Neural Network Accelerator With Efficient Index Matching. IEEE Open Journal of Circuits and Systems, 2021, 2, 144-155.	1.4	1
13	Mixed-Signal Computing for Deep Neural Network Inference. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 2021, 29, 3-13.	2.1	46
14	FantastIC4: A Hardware-Software Co-Design Approach for Efficiently Running 4Bit-Compact Multilayer Perceptrons. IEEE Open Journal of Circuits and Systems, 2021, 2, 407-419.	1.4	7
15	Web Based GPU Acceleration in Embodied Agent Training Workflow. , 2021, , .		0
16	Learning Sparse Neural Networks Using Non-Convex Regularization. IEEE Transactions on Emerging Topics in Computational Intelligence, 2022, 6, 287-299.	3.4	6
17	Compression strategies and space-conscious representations for deep neural networks. , 2021, , .		4
18	Design and Implementation of Deep Learning Based Contactless Authentication System Using Hand Gestures. Electronics (Switzerland), 2021, 10, 182.	1.8	32

#	Article	IF	Citations
19	Custom Hardware Architectures for Deep Learning on Portable Devices: A Review. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33, 6068-6088.	7.2	21
20	Towards Model Compression for Deep Learning Based Speech Enhancement. IEEE/ACM Transactions on Audio Speech and Language Processing, 2021, 29, 1785-1794.	4.0	40
21	Kronecker CP Decomposition With Fast Multiplication for Compressing RNNs. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34, 2205-2219.	7.2	3
22	Contactless Human Monitoring: Challenges and Future Direction. Intelligent Systems Reference Library, 2021, , 335-364.	1.0	2
24	Dep-$$L_0$$: Improving $$L_0$$-Based Network Sparsification via Dependency Modeling. Lecture Notes in Computer Science, 2021, , 167-183.	1.0	0
25	DNN Model Compression for IoT Domain-Specific Hardware Accelerators. IEEE Internet of Things Journal, 2022, 9, 6650-6662.	5.5	13
26	Efficient Environmental Context Prediction for Lower Limb Prostheses. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2022, 52, 3980-3994.	5.9	17
27	Compute-in-Memory Chips for Deep Learning: Recent Trends and Prospects. IEEE Circuits and Systems Magazine, 2021, 21, 31-56.	2.6	115
28	Digital Retina: A Way to Make the City Brain More Efficient by Visual Coding. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31, 4147-4161.	5.6	19
29	Hardware accelerator for training with integer backpropagation and probabilistic weight update. Advances in Computers, 2021, , 343-365.	1.2	3
30	A Comprehensive Survey on Training Acceleration for Large Machine Learning Models in IoT. IEEE Internet of Things Journal, 2022, 9, 939-963.	5.5	14
31	Modeling and Optimization of SRAM-based In-Memory Computing Hardware Design. , 2021, , .		9
32	A FeRAM based Volatile/Non-volatile Dual-mode Buffer Memory for Deep Neural Network Training. , 2021, , .		2
33	A Runtime Reconfigurable Design of Compute-in-Memory based Hardware Accelerator. , 2021, , .		3
34	Characterization and Mitigation of Relaxation Effects on Multi-level RRAM based In-Memory Computing. , 2021, , .		6
35	Lane Compression. Transactions on Embedded Computing Systems, 2021, 20, 1-26.	2.1	1
36	Carry-Propagation-Adder-Factored Gemmini Systolic Array for Machine Learning Acceleration. Electronics (Switzerland), 2021, 10, 652.	1.8	4
37	Regularization-Free Structural Pruning for GPU Inference Acceleration. , 2021, , .		0

#	Article	IF	Citations
38	Improving DNN Fault Tolerance using Weight Pruning and Differential Crossbar Mapping for ReRAM-based Edge AI. , 2021, , .		19
39	APAE: an IoT intrusion detection system using asymmetric parallel auto-encoder. Neural Computing and Applications, 2023, 35, 4813-4833.	3.2	18
40	Accelerating deep neural networks for efficient scene understanding in automotive cyber-physical systems. , 2021, , .		8
41	HSC: A Hybrid Spin/CMOS Logic Based In-Memory Engine with Area-Efficient Mapping Strategy. , 2021, , .		0
42	Deep Learning-Based Sign Language Digits Recognition From Thermal Images With Edge Computing System. IEEE Sensors Journal, 2021, 21, 10445-10453.	2.4	33
43	Simplified Hardware Implementation of Memoryless Dot Product for Neural Network Inference. , 2021, , .		3
44	Overparametrization of HyperNetworks at Fixed FLOP-Count Enables Fast Neural Image Enhancement. , 2021, , .		2
45	Advancements in Microprocessor Architecture for Ubiquitous AIâ€"An Overview on History, Evolution, and Upcoming Challenges in AI Implementation. Micromachines, 2021, 12, 665.	1.4	11
46	Compute-in-RRAM with Limited On-chip Resources. , 2021, , .		1
47	A Runtime Reconfigurable Design of Compute-in-Memoryâ€"Based Hardware Accelerator for Deep Learning Inference. ACM Transactions on Design Automation of Electronic Systems, 2021, 26, 1-18.	1.9	4
48	NeuroSim Validation with 40nm RRAM Compute-in-Memory Macro. , 2021, , .		7
49	NeuroSim Simulator for Compute-in-Memory Hardware Accelerator: Validation and Benchmark. Frontiers in Artificial Intelligence, 2021, 4, 659060.	2.0	23
50	SNPE-SRGAN: Lightweight Generative Adversarial Networks for Single-Image Super-Resolution on Mobile Using SNPE Framework. Journal of Physics: Conference Series, 2021, 1898, 012038.	0.3	1
51	Towards Inference Delivery Networks: Distributing Machine Learning with Optimality Guarantees. , 2021, , .		7
52	Progressive principle component analysis for compressing deep convolutional neural networks. Neurocomputing, 2021, 440, 197-206.	3.5	7
53	Compressing Deep Neural Networks for Efficient Speech Enhancement. , 2021, , .		5
54	Towards Automatic and Agile AI/ML Accelerator Design with End-to-End Synthesis. , 2021, , .		4
55	FA-GAL-ResNet: Lightweight Residual Network using Focused Attention Mechanism and Generative Adversarial Learning via Knowledge Distillation. , 2021, , .		1

#	Article	IF	Citations
56	Privacy-Preserving Deep Learning Based on Multiparty Secure Computation: A Survey. IEEE Internet of Things Journal, 2021, 8, 10412-10429.	5.5	8
57	Hardware Aspects of Parallel Neural Network Implementation. , 2021, , .		0
58	Ps and Qs: Quantization-Aware Pruning for Efficient Low Latency Neural Network Inference. Frontiers in Artificial Intelligence, 2021, 4, 676564.	2.0	15
59	A Marr's Threeâ€Level Analytical Framework for Neuromorphic Electronic Systems. Advanced Intelligent Systems, 2021, 3, 2100054.	3.3	3
60	Intragroup sparsity for efficient inference. , 2021, , .		0
61	CARLA: A Convolution Accelerator With a Reconfigurable and Low-Energy Architecture. IEEE Transactions on Circuits and Systems I: Regular Papers, 2021, 68, 3184-3196.	3.5	7
62	Optimization of Online Education and Teaching Evaluation System Based on GA-BP Neural Network. Computational Intelligence and Neuroscience, 2021, 2021, 1-9.	1.1	13
63	Compacting Deep Neural Networks for Internet of Things: Methods and Applications. IEEE Internet of Things Journal, 2021, 8, 11935-11959.	5.5	27
64	AILC: Accelerate On-Chip Incremental Learning With Compute-in-Memory Technology. IEEE Transactions on Computers, 2021, 70, 1225-1238.	2.4	9
65	Hardware-Aware Neural Architecture Search: Survey and Taxonomy. , 2021, , .		41
66	Pruning-Aware Merging for Efficient Multitask Inference. , 2021, , .		1
67	QTTNet: Quantized tensor train neural networks for 3D object and video recognition. Neural Networks, 2021, 141, 420-432.	3.3	16
68	Zero-Shot Learning Of A Conditional Generative Adversarial Network For Data-Free Network Quantization. , 2021, , .		1
69	Urban Fine Management of Multisource Spatial Data Fusion Based on Smart City Construction. Mathematical Problems in Engineering, 2021, 2021, 1-10.	0.6	3
70	GenExp: Multi-objective pruning for deep neural network based on genetic algorithm. Neurocomputing, 2021, 451, 81-94.	3.5	19
71	Landscape Planning and Image Analysis Based on Multipopulation Coevolution Particle Swarm Radial Basis Function Neural Network Algorithm. Computational Intelligence and Neuroscience, 2021, 2021, 1-11.	1.1	1
72	A Low-Cost, Low-Power and Real-Time Image Detector for Grape Leaf Esca Disease Based on a Compressed CNN. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 2021, 11, 468-481.	2.7	16
73	3D Virtual Reality Implementation of Tourist Attractions Based on the Deep Belief Neural Network. Computational Intelligence and Neuroscience, 2021, 2021, 1-11.	1.1	2

#	Article	IF	Citations
74	A Novel Ultra-Low Power 8T SRAM-Based Compute-in-Memory Design for Binary Neural Networks. Electronics (Switzerland), 2021, 10, 2181.	1.8	3
75	Application of Deep Reinforcement Learning Algorithm in Uncertain Logistics Transportation Scheduling. Computational Intelligence and Neuroscience, 2021, 2021, 1-9.	1.1	2
76	Learnable Heterogeneous Convolution: Learning both topology and strength. Neural Networks, 2021, 141, 270-280.	3.3	1
77	Mobility- and Energy-Aware Cooperative Edge Offloading for Dependent Computation Tasks. Network, 2021, 1, 191-214.	1.5	14
78	Hardware Acceleration of Sparse and Irregular Tensor Computations of ML Models: A Survey and Insights. Proceedings of the IEEE, 2021, 109, 1706-1752.	16.4	35
79	Nonlinear tensor train format for deep neural network compression. Neural Networks, 2021, 144, 320-333.	3.3	14
80	BISWSRBS: A Winograd-based CNN Accelerator with a Fine-grained Regular Sparsity Pattern and Mixed Precision Quantization. ACM Transactions on Reconfigurable Technology and Systems, 2021, 14, 1-28.	1.9	1
81	Model compression for on-device inference. , 2022, , 71-82.		1
82	Spartan: A Sparsity-Adaptive Framework to Accelerate Deep Neural Network Training on GPUs. IEEE Transactions on Parallel and Distributed Systems, 2021, 32, 2448-2463.	4.0	5
83	Hybrid neural state machine for neural network. Science China Information Sciences, 2021, 64, 1.	2.7	7
84	Enabling and Leveraging AI in the Intelligent Edge: A Review of Current Trends and Future Directions. IEEE Open Journal of the Communications Society, 2021, 2, 2311-2341.	4.4	2
85	A Lightweight Deep Learning Algorithm for WiFi-Based Identity Recognition. IEEE Internet of Things Journal, 2021, 8, 17449-17459.	5.5	10
86	Intelligent Radio Signal Processing: A Survey. IEEE Access, 2021, 9, 83818-83850.	2.6	49
87	IdleSR: Efficient Super-Resolution Network with Multi-scale IdleBlocks. Lecture Notes in Computer Science, 2020, , 136-151.	1.0	2
88	A Proximal Iteratively Reweighted Approach for Efficient Network Sparsification. IEEE Transactions on Computers, 2022, 71, 185-196.	2.4	1
89	Light-YOLOv4: An Edge-Device Oriented Target Detection Method for Remote Sensing Images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14, 10808-10820.	2.3	34
90	Automatic Modulation Classification: A Deep Architecture Survey. IEEE Access, 2021, 9, 142950-142971.	2.6	50
91	Secure XOR-CIM Engine: Compute-In-Memory SRAM Architecture With Embedded XOR Encryption. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 2021, 29, 2027-2039.	2.1	8

#	Article	IF	Citations
92	SpecMCTS: Accelerating Monte Carlo Tree Search Using Speculative Tree Traversal. IEEE Access, 2021, 9, 142195-142205.	2.6	1
93	An FPGA-based MobileNet Accelerator Considering Network Structure Characteristics. , 2021, , .		14
95	International Trade Path with Multi-Polarization based on Multidirectional Mutation Genetic Algorithm Enabled Neural Network. Computational Intelligence and Neuroscience, 2021, 2021, 1-9.	1.1	1
96	On-Device Object Detection for More Efficient and Privacy-Compliant Visual Perception in Context-Aware Systems. Applied Sciences (Switzerland), 2021, 11, 9173.	1.3	2
97	Pruning Meta-Trained Networks for On-Device Adaptation. , 2021, , .		5
98	PARAFAC2 and local minima. Chemometrics and Intelligent Laboratory Systems, 2021, 219, 104446.	1.8	7
99	Real-Time Instance Segmentation for Low-Cost Mobile Robot Systems Based on Computation Offloading. , 2021, , .		1
100	A New Clustering-Based Technique for the Acceleration of Deep Convolutional Networks. , 2020, , .		3
101	A Ferroelectric-based Volatile/Non-volatile Dual-mode Buffer Memory for Deep Neural Network Accelerators. IEEE Transactions on Computers, 2021, , 1-1.	2.4	2
102	Edge Intelligence for Smart Metro Systems: Architecture and Enabling Technologies. IEEE Network, 2022, 36, 136-143.	4.9	3
103	SecureTrain: An Approximation-Free and Computationally Efficient Framework for Privacy-Preserved Neural Network Training. IEEE Transactions on Network Science and Engineering, 2022, 9, 187-202.	4.1	5
104	Abstract Neural Networks. Lecture Notes in Computer Science, 2020, , 65-88.	1.0	11
105	On-Device Deep Multi-Task Inference via Multi-Task Zipping. IEEE Transactions on Mobile Computing, 2023, 22, 2878-2891.	3.9	1
106	Zero-shot Adversarial Quantization. , 2021, , .		37
107	Bringing AI to edge: From deep learningâ€™s perspective. Neurocomputing, 2022, 485, 297-320.	3.5	44
108	Leveraging Noise and Aggressive Quantization of In-Memory Computing for Robust DNN Hardware Against Adversarial Input and Weight Attacks. , 2021, , .		4
109	XOR-CIM. , 2020, , .		13
110	Heterogeneous Systolic Array Architecture for Compact CNNs Hardware Accelerators. IEEE Transactions on Parallel and Distributed Systems, 2021, , 1-1.	4.0	3

#	Article	IF	Citations
111	Cross-layer knowledge distillation with KL divergence and offline ensemble for compressing deep neural network. APSIPA Transactions on Signal and Information Processing, 2021, 10, .	2.6	1
112	Photon-Driven Neural Reconstruction for Path Guiding. ACM Transactions on Graphics, 2022, 41, 1-15.	4.9	4
113	Forest Fire Detection Based on Lightweight Yolo. , 2021, , .		10
114	Accelerating 3D scene analysis for autonomous driving on embedded AI computing platforms. , 2021, , .		2
115	A Scatter-and-Gather Spiking Convolutional Neural Network on a Reconfigurable Neuromorphic Hardware. Frontiers in Neuroscience, 2021, 15, 694170.	1.4	4
116	A New Clustering-Based Technique for the Acceleration of Deep Convolutional Networks. Advances in Intelligent Systems and Computing, 2022, , 123-150.	0.5	2
117	MASS. , 2021, , .		1
119	Neural Architecture Search and Hardware Accelerator Co-Search: A Survey. IEEE Access, 2021, 9, 151337-151362.	2.6	19
120	Optimal hybrid heat transfer search and grey wolf optimization-based homomorphic encryption model to assure security in cloud-based IoT environment. Peer-to-Peer Networking and Applications, 2022, 15, 703-723.	2.6	6
121	Compression of Deep Learning Models for Text: A Survey. ACM Transactions on Knowledge Discovery From Data, 2022, 16, 1-55.	2.5	23
122	Sparse CapsNet with explicit regularizer. Pattern Recognition, 2022, 124, 108486.	5.1	5
123	Mixture of Deterministic and Stochastic Quantization Schemes for Lightweight CNN. , 2020, , .		4
124	Low-Complexity Recurrent Neural Network Based Equalizer With Embedded Parallelization for 100-Gbit/s/Î» PON. Journal of Lightwave Technology, 2022, 40, 1353-1359.	2.7	13
126	Multimedia Data Analysis With Edge Computing. IEEE MultiMedia, 2021, 28, 5-7.	1.5	2
127	SME: ReRAM-based Sparse-Multiplication-Engine to Squeeze-Out Bit Sparsity of Neural Network. , 2021, , .		11
128	Compression of Time Series Classification Model MC-MHLF using Knowledge Distillation. , 2021, , .		0
129	Pre-RTL DNN Hardware Evaluator With Fused Layer Support. , 2021, , .		1
130	Efficient methods for deep learning. , 2022, , 159-190.		5

#	Article	IF	Citations
131	A Review of Efficient Real-Time Decision Making in the Internet of Things. Technologies, 2022, 10, 12.	3.0	3
132	Multimodal Neural Network Acceleration on a Hybrid CPU-FPGA Architecture: A Case Study. IEEE Access, 2022, 10, 9603-9617.	2.6	2
134	DFE: efficient IoT network intrusion detection using deep feature extraction. Neural Computing and Applications, 2022, 34, 15175-15195.	3.2	10
135	Machine Learning for Multimedia Communications. Sensors, 2022, 22, 819.	2.1	4
136	AntiDoteX: Attention-Based Dynamic Optimization for Neural Network Runtime Efficiency. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2022, 41, 4694-4707.	1.9	0
137	Structured precision skipping: Accelerating convolutional neural networks with budget-aware dynamic precision selection. Journal of Systems Architecture, 2022, 124, 102403.	2.5	2
138	Accelerating On-Chip Training with Ferroelectric-Based Hybrid Precision Synapse. ACM Journal on Emerging Technologies in Computing Systems, 2022, 18, 1-20.	1.8	1
139	Cellular, Wide-Area, and Non-Terrestrial IoT: A Survey on 5G Advances and the Road Toward 6G. IEEE Communications Surveys and Tutorials, 2022, 24, 1117-1174.	24.8	172
140	Homecare-Oriented ECG Diagnosis With Large-Scale Deep Neural Network for Continuous Monitoring on Embedded Devices. IEEE Transactions on Instrumentation and Measurement, 2022, 71, 1-13.	2.4	20
141	Self-supervised learning for medieval handwriting identification: A case study from the Vatican Apostolic Library. Information Processing and Management, 2022, 59, 102875.	5.4	5
142	KeepEdge: A Knowledge Distillation Empowered Edge Intelligence Framework for Visual Assisted Positioning in UAV Delivery. IEEE Transactions on Mobile Computing, 2023, 22, 4729-4741.	3.9	5
143	A Technique for Approximate Communication in Network-on-Chips for Image Classification. IEEE Transactions on Emerging Topics in Computing, 2023, 11, 30-42.	3.2	2
144	IVQ: In-Memory Acceleration of DNN Inference Exploiting Varied Quantization. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2022, 41, 5313-5326.	1.9	2
145	Skydiver: A Spiking Neural Network Accelerator Exploiting Spatio-Temporal Workload Balance. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2022, 41, 5732-5736.	1.9	12
146	A Robust Authentication and Authorization System Powered by Deep Learning and Incorporating Hand Signals. Lecture Notes on Data Engineering and Communications Technologies, 2022, , 1061-1071.	0.5	3
147	Neural Network Structure Optimization by Simulated Annealing. Entropy, 2022, 24, 348.	1.1	12
148	Performance Modeling of Computer Vision-based CNN on Edge GPUs. Transactions on Embedded Computing Systems, 2022, 21, 1-33.	2.1	1
149	A Construction Kit for Efficient Low Power Neural Network Accelerator Designs. Transactions on Embedded Computing Systems, 2022, 21, 1-36.	2.1	0

#	Article	IF	Citations
150	Weight-Quantized SqueezeNet for Resource-Constrained Robot Vacuums for Indoor Obstacle Classification. AI, 2022, 3, 180-193.	2.1	28
151	Toward Open-World Electroencephalogram Decoding Via Deep Learning: A comprehensive survey. IEEE Signal Processing Magazine, 2022, 39, 117-134.	4.6	37
152	The Bitlet Model: A Parameterized Analytical Model to Compare PIM and CPU Systems. ACM Journal on Emerging Technologies in Computing Systems, 2022, 18, 1-29.	1.8	10
153	Low-power deep learning edge computing platform for resource constrained lightweight compact UAVs. Sustainable Computing: Informatics and Systems, 2022, 34, 100725.	1.6	7
154	Algorithm/architecture solutions to improve beyond uniform quantization in embedded DNN accelerators. Journal of Systems Architecture, 2022, 126, 102454.	2.5	0
155	Enable Deep Learning on Mobile Devices: Methods, Systems, and Applications. ACM Transactions on Design Automation of Electronic Systems, 2022, 27, 1-50.	1.9	38
156	PDAE: Efficient network intrusion detection in IoT using parallel deep auto-encoders. Information Sciences, 2022, 598, 57-74.	4.0	27
157	Golden subject is everyone: A subject transfer neural network for motor imagery-based brain computer interfaces. Neural Networks, 2022, 151, 111-120.	3.3	25
158	Towards Memory-Efficient Neural Networks via Multi-Level in situ Generation. , 2021, , .		1
159	A data-aware dictionary-learning based technique for the acceleration of deep convolutional networks. , 2021, , .		1
160	Improving Neural Network Efficiency via Post-training Quantization with Adaptive Floating-Point. , 2021, , .		22
161	Towards Mixed-Precision Quantization of Neural Networks via Constrained Optimization. , 2021, , .		18
162	Chebyshev Polynomial Broad Learning System. , 2021, , .		1
163	Deep Learning Approach at the Edge to Detect Iron Ore Type. Sensors, 2022, 22, 169.	2.1	3
164	Optimized convolutional neural network architectures for efficient on-device vision-based object detection. Neural Computing and Applications, 2022, 34, 10469-10501.	3.2	10
165	A New Method to Compress Neural Networks. , 2021, , .		0
166	Permutation-Invariant Representation ofÂNeural Networks withÂNeuron Embeddings. Lecture Notes in Computer Science, 2022, , 294-308.	1.0	1
167	QLP: Deep Q-Learning for Pruning Deep Neural Networks. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32, 6488-6501.	5.6	9

#	Article	IF	Citations
168	ECQ$$^{ext {x}}$$: Explainability-Driven Quantization forÂLow-Bit andÂSparse DNNs. Lecture Notes in Computer Science, 2022, , 271-296.	1.0	5
169	Minimum signed digit approximation for faster and more efficient convolutional neural network computation on embedded devices. Engineering Science and Technology, an International Journal, 2022, 36, 101153.	2.0	5
171	Sensors in Hospitals. , 2022, , .		0
172	SoBS-X: Squeeze-Out Bit Sparsity for ReRAM-Crossbar-Based Neural Network Accelerator. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2023, 42, 204-217.	1.9	2
173	A Low-Rank CNN Architecture for Real-Time Semantic Segmentation in Visual SLAM Applications. IEEE Open Journal of Circuits and Systems, 2022, 3, 115-133.	1.4	5
174	Towards privacy aware deep learning for embedded systems. , 2022, , .		2
175	åŸºäºŽè¿›åŒ–ç–•ç‚„è‡ªé€‚åº”è•é‚¦å¯¦çŽ°æ³•. Scientia Sinica Informationis, 2022, , .	0.2	0
176	Signal Compression via Neural Implicit Representations. , 2022, , .		1
177	Embedded Edge Artificial Intelligence for Longitudinal Rip Detection in Conveyor Belt Applied at the Industrial Mining Environment. SN Computer Science, 2022, 3, 1.	2.3	4
178	Stochastic batch size for adaptive regularization in deep network optimization. Pattern Recognition, 2022, 129, 108776.	5.1	2
179	Griffin: Rethinking Sparse Optimization for Deep Learning Architectures. , 2022, , .		3
180	THETA: A High-Efficiency Training Accelerator for DNNs With Triple-Side Sparsity Exploration. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 2022, 30, 1034-1046.	2.1	5
181	BMPQ: Bit-Gradient Sensitivity-Driven Mixed-Precision Quantization of DNNs from Scratch. , 2022, , .		6
182	BenQ: Benchmarking Automated Quantization on Deep Neural Network Accelerators. , 2022, , .		1
183	Spiking Neural Network Integrated Circuits: A Review of Trends and Future Directions. , 2022, , .		28
184	Anchor pruning for object detection. Computer Vision and Image Understanding, 2022, 221, 103445.	3.0	4
185	SplitPlace: AI Augmented Splitting and Placement of Large-Scale Neural Networks in Mobile Edge Environments. IEEE Transactions on Mobile Computing, 2023, 22, 5539-5554.	3.9	8
186	A survey on hardware accelerators: Taxonomy, trends, challenges, and perspectives. Journal of Systems Architecture, 2022, 129, 102561.	2.5	27

#	Article	IF	Citations
187	LNNet: Lightweight Nested Network for motion deblurring. Journal of Systems Architecture, 2022, , 102584.	2.5	2
188	Data Stream Oriented Fine-grained Sparse CNN Accelerator with Efficient Unstructured Pruning Strategy. , 2022, , .		3
189	Integrating deep learning and rule-based systems into a smart devices decision support system for visual inspection in production. Procedia CIRP, 2022, 109, 305-310.	1.0	3
190	Camel: Managing Data for Efficient Stream Learning. , 2022, , .		4
191	Efficient neural network representations for energy data analytics on embedded systems. , 2022, , .		2
192	Distributed Edge System Orchestration for Web-based Mobile Augmented Reality Services. IEEE Transactions on Services Computing, 2022, , 1-15.	3.2	6
193	Stage-Wise Magnitude-Based Pruning for Recurrent Neural Networks. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35, 1666-1680.	7.2	2
194	A Path Relinking Method for the Joint Online Scheduling and Capacity Allocation of DL Training Workloads in GPU as a Service Systems. IEEE Transactions on Services Computing, 2022, , 1-16.	3.2	0
195	LRP-based Policy Pruning and Distillation of Reinforcement Learning Agents for Embedded Systems. , 2022, , .		4
196	Robust 4D awareness via diffusion adaptation over Connected and Autonomated vehicles. , 2022, , .		0
197	Optical processor for a binarized neural network. Optics Letters, 2022, 47, 3892.	1.7	6
198	Only-train-electrical-to-optical-conversion (OTEOC): simple diffractive neural networks with optical readout. Optics Express, 2022, 30, 28024.	1.7	4
199	abstractPIM: Bridging the Gap Between Processing-In-Memory Technology and Instruction Set Architecture. , 2020, , .		4
200	Optimization Tools for ConvNets on the Edge. , 2020, , .		0
201	Digital Versus Analog Artificial Intelligence Accelerators: Advances, trends, and emerging designs. IEEE Solid-State Circuits Magazine, 2022, 14, 65-79.	0.5	18
202	Winograd convolution. , 2022, , .		8
203	Structured Dynamic Precision for Deep Neural Networks Quantization. ACM Transactions on Design Automation of Electronic Systems, 2023, 28, 1-24.	1.9	0
204	Joint Architecture Design and Workload Partitioning for DNN Inference on Industrial IoT Clusters. ACM Transactions on Internet Technology, 2023, 23, 1-21.	3.0	1

#	Article	IF	Citations
205	Weak self-supervised learning for seizure forecasting: a feasibility study. Royal Society Open Science, 2022, 9, .	1.1	8
206	Quantization and sparsity-aware processing for energy-efficient NVM-based convolutional neural networks. Frontiers in Electronics, 0, 3, .	2.0	1
207	Towards efficient full 8-bit integer DNN online training on resource-limited devices without batch normalization. Neurocomputing, 2022, 511, 175-186.	3.5	1
208	Multiâ€objective evolutionary optimization for hardwareâ€aware neural network pruning. Fundamental Research, 2022, , .	1.6	7
209	p-Meta. , 2022, , .		1
210	Efficient Visual Recognition: A Survey on Recent Advances and Brain-inspired Methodologies. , 2022, 19, 366-411.		9
212	Realistic acceleration of neural networks with fine-grained tensor decomposition. Neurocomputing, 2022, 512, 52-68.	3.5	2
213	Multiuser Co-Inference With Batch Processing Capable Edge Server. IEEE Transactions on Wireless Communications, 2023, 22, 286-300.	6.1	3
214	A Survey of Intelligent Chip Design Research Based on Spiking Neural Networks. IEEE Access, 2022, 10, 89663-89686.	2.6	3
215	FACCU: Enable Fast Accumulation for High-Speed DSP Systems. IEEE Transactions on Circuits and Systems II: Express Briefs, 2022, 69, 4634-4638.	2.2	1
216	Real Time Power Equipment Meter Recognition Based on Deep Learning. IEEE Transactions on Instrumentation and Measurement, 2022, 71, 1-15.	2.4	12
217	Machine and Deep Learning for Resource Allocation in Multi-Access Edge Computing: A Survey. IEEE Communications Surveys and Tutorials, 2022, 24, 2449-2494.	24.8	19
218	Edge AI: Leveraging the Full Potential of Deep Learning. Studies in Computational Intelligence, 2022, , 27-46.	0.7	5
219	Feature Distillation Siamese Networks for Object Tracking. SSRN Electronic Journal, 0, , .	0.4	0
220	Hardware-friendly compression and hardware acceleration for transformer: A survey. Electronic Research Archive, 2022, 30, 3755-3785.	0.4	1
221	Edge-Assisted Real-Time Instance Segmentation for Resource-Limited IoT Devices. IEEE Internet of Things Journal, 2023, 10, 473-485.	5.5	2
222	Pruning for Compression of Visual Pattern Recognition Networks: A Survey from Deep Neural Networks Perspective. Lecture Notes in Electrical Engineering, 2022, , 675-687.	0.3	4
223	Compressing Models with Few Samples: Mimicking then Replacing. , 2022, , .		4

#	Article	IF	Citations
224	Approximate Bisimulation Relations for Neural Networks and Application to Assured Neural Network Compression. , 2022, , .		2
225	AdaSTE: An Adaptive Straight-Through Estimator to Train Binary Neural Networks. , 2022, , .		3
226	DiSparse: Disentangled Sparsification for Multitask Model Compression. , 2022, , .		4
227	DualPIM: A Dual-Precision and Low-Power CNN Inference Engine Using SRAM- and eDRAM-based Processing-in-Memory Arrays. , 2022, , .		1
228	LTH-ECG: Lottery Ticket Hypothesis-based Deep Learning Model Compression for Atrial Fibrillation Detection from Single Lead ECG On Wearable and Implantable Devices. , 2022, , .		1
229	MCA-YOLOV5-Light: A Faster, Stronger and Lighter Algorithm for Helmet-Wearing Detection. Applied Sciences (Switzerland), 2022, 12, 9697.	1.3	9
230	A Method of Deep Learning Model Optimization for Image Classification on Edge Device. Sensors, 2022, 22, 7344.	2.1	4
231	FedQNN: A Computationâ€“Communication-Efficient Federated Learning Framework for IoT With Low-Bitwidth Neural Network Quantization. IEEE Internet of Things Journal, 2023, 10, 2494-2507.	5.5	2
232	Forming-free titanium oxide neuromorphic crossbar array for robotics and AI systems. , 2022, , .		0
233	Approximate Network-on-Chips with Application to Image Classification. , 2022, , .		0
234	Neuron Specific Pruning for Communication Efficient Federated Learning. , 2022, , .		1
235	Halftoning with Multi-Agent Deep Reinforcement Learning. , 2022, , .		3
236	ReLP: Reinforcement Learning Pruning Method Based on Prior Knowledge. Neural Processing Letters, 2023, 55, 4661-4678.	2.0	1
237	Partitioning DNNs for Optimizing Distributed Inference Performance on Cooperative Edge Devices: A Genetic Algorithm Approach. Applied Sciences (Switzerland), 2022, 12, 10619.	1.3	3
238	An Offloading Algorithm for Maximizing Inference Accuracy on Edge Device in an Edge Intelligence System. , 2022, , .		2
239	A Novel Tran_NAS Method for the Identification of Fe- and Mg-Deficient Pear Leaves from N- and P-Deficient Pear Leaf Data. ACS Omega, 2022, 7, 39727-39741.	1.6	4
240	LRPâ€based network pruning and policy distillation of robust and nonâ€robust DRL agents for embedded systems. Concurrency Computation Practice and Experience, 2023, 35, .	1.4	3
241	Combining Non-Data-Adaptive Transforms for OCT Image Denoising by Iterative Basis Pursuit. , 2022, , .		3

#	Article	IF	Citations
242	Artificial Tactile Recognition Enabled by Flexible Low-Voltage Organic Transistors and Low-Power Synaptic Electronics. ACS Applied Materials & Interfaces, 2022, 14, 48948-48959.	4.0	15
243	Multiple hierarchical compression for deep neural network toward intelligent bearing fault diagnosis. Engineering Applications of Artificial Intelligence, 2022, 116, 105498.	4.3	20
244	High performance inference of gait recognition models on embedded systems. Sustainable Computing: Informatics and Systems, 2022, 36, 100814.	1.6	1
245	Fast visual tracking with lightweight Siamese network and template-guided learning. Knowledge-Based Systems, 2022, 258, 110037.	4.0	1
246	Distributed Artificial Intelligence Empowered by End-Edge-Cloud Computing: A Survey. IEEE Communications Surveys and Tutorials, 2023, 25, 591-624.	24.8	34
247	Variance-Guided Structured Sparsity in Deep Neural Networks. IEEE Transactions on Artificial Intelligence, 2023, 4, 1714-1723.	3.4	0
248	Hardware Implementation of Stochastic Computing-based Morphological Neural Systems. , 2022, , .		3
249	Redundancy Pruning for Binary Hyperdimensional Computing Architectures. , 2022, , .		1
250	Transforming Large-Size to Lightweight Deep Neural Networks for IoT Applications. ACM Computing Surveys, 2023, 55, 1-35.	16.1	7
251	Traffic prediction using artificial intelligence: Review of recentÂadvances and emerging opportunities. Transportation Research Part C: Emerging Technologies, 2022, 145, 103921.	3.9	26
252	Lightweight Deep Learning Model for Radar-Based Fall Detection With Metric Learning. IEEE Internet of Things Journal, 2023, 10, 8111-8122.	5.5	3
253	Feature distillation Siamese networks for object tracking. Applied Soft Computing Journal, 2023, 132, 109912.	4.1	3
254	Beyond Transmitting Bits: Context, Semantics, and Task-Oriented Communications. IEEE Journal on Selected Areas in Communications, 2023, 41, 5-41.	9.7	66
255	Hardware-aware neural architecture search for stochastic computing-based neural networks on tiny devices. Journal of Systems Architecture, 2023, 135, 102810.	2.5	1
256	Deep neural networks compression: A comparative survey and choice recommendations. Neurocomputing, 2023, 520, 152-170.	3.5	26
257	Adversarial learning-based multi-level dense-transmission knowledge distillation for AP-ROP detection. Medical Image Analysis, 2023, 84, 102725.	7.0	2
258	Efficient Acceleration of Deep Learning Inference on Resource-Constrained Edge Devices: A Review. Proceedings of the IEEE, 2023, 111, 42-91.	16.4	18
259	SMOF: Squeezing More Out ofÂFilters Yields Hardware-Friendly CNN Pruning. Lecture Notes in Computer Science, 2022, , 242-254.	1.0	1

#	Article	IF	Citations
260	Reconfigurability, Why It Matters in AI Tasks Processing: A Survey of Reconfigurable AI Chips. IEEE Transactions on Circuits and Systems I: Regular Papers, 2023, 70, 1228-1241.	3.5	1
261	Kernel Modulation: A Parameter-Efficient Method for Training Convolutional Neural Networks. , 2022, , .		0
262	Light-Weight EPINET Architecture for Fast Light Field Disparity Estimation. , 2022, , .		2
263	Comparing the performance of multi-layer perceptron training on electrical and optical network-on-chips. Journal of Supercomputing, 2023, 79, 10725-10746.	2.4	3
264	LightMobileNetV2: A Lightweight Model for the Classification of COVID-19 Using Chest X-Ray Images. Lecture Notes in Networks and Systems, 2023, , 142-150.	0.5	0
265	A Novel Approach to Structured Pruning of Neural Network for Designing Compact Audio-Visual Wake Word Spotting System. , 2022, , .		0
266	Hardware/Software Co-acceleration of Progressive Learning under Feature Dimension Variation. , 2022, , .		0
267	é¢åʻäº‹»¶ç›¸æœºçš„è½»é‡åŒ–è„‰å†²è¯†åˆ«ç½‘ç»œ. Scientia Sinica Informationis, 2022, , .	0.2	0
268	A 1.6-mW Sparse Deep Learning Accelerator for Speech Separation. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 2023, 31, 310-319.	2.1	0
269	An Efficient Lightweight Event Detection Algorithm for On-Site Non-Intrusive Load Monitoring. IEEE Transactions on Instrumentation and Measurement, 2023, 72, 1-13.	2.4	4
270	Reducing Computational Complexity of Neural Networks in Optical Channel Equalization: From Concepts to Implementation. Journal of Lightwave Technology, 2023, 41, 4557-4581.	2.7	7
271	Model Compression for Non-Sequential and Sequential Visual Pattern Recognition Networks â€• A Hybrid Approach. , 2022, , .		0
272	Safety and Performance, Why not Both? Bi-Objective Optimized Model Compression toward AI Software Deployment. , 2022, , .		1
273	Magnitude and Similarity Based Variable Rate Filter Pruning for Efficient Convolution Neural Networks. Applied Sciences (Switzerland), 2023, 13, 316.	1.3	2
274	Safety Verification of Neural Network Control Systems Using Guaranteed Neural Network Model Reduction. , 2022, , .		0
275	Testability and Dependability of AI Hardware: Survey, Trends, Challenges, and Perspectives. IEEE Design and Test, 2023, 40, 8-58.	1.1	8
276	Explainable Network Pruning for Model Acceleration Based on Filter Similarity and Importance. Lecture Notes in Computer Science, 2023, , 214-229.	1.0	0
277	Learning Lightweight Neural Networks via Channel-Split Recurrent Convolution. , 2023, , .		0

#	Article	IF	Citations
278	Hardware-Software Codesign of DNN Accelerators Using Approximate Posit Multipliers. , 2023, , .		1
279	Multi-Part Knowledge Distillation for the Efficient Classification of Colorectal Cancer Histology Images. , 2022, , .		1
280	A Serverless Computing Fabric for Edge & Cloud. , 2022, , .		8
281	Neural-Network-Assisted Packet Accelerators for Internet of Things Network Systems. IEEE Internet of Things Journal, 2023, 10, 15238-15251.	5.5	1
282	A unifying review of edge intelligent computing technique applications in the field of energy networks. Journal of Industrial and Management Optimization, 2023, 19, 7966-7992.	0.8	2
283	Fast data-free model compression via dictionary-pair reconstruction. Knowledge and Information Systems, 0, , .	2.1	0
284	Diluted binary neural network. Pattern Recognition, 2023, 140, 109556.	5.1	1
285	TailorFL. , 2022, , .		3
286	An effective low-rank compression with a joint rank selection followed by a compression-friendly training. Neural Networks, 2023, 161, 165-177.	3.3	4
287	Learning broad learning system with controllable sparsity through L0 regularization. Applied Soft Computing Journal, 2023, 136, 110068.	4.1	0
288	Horizontally Distributed Inference of Deep Neural Networks for AI-Enabled IoT. Sensors, 2023, 23, 1911.	2.1	4
289	Research Challenges, Recent Advances, and Popular Datasets in Deep Learning-Based Underwater Marine Object Detection: A Review. Sensors, 2023, 23, 1990.	2.1	7
290	Neuromorphic processor-oriented hybrid Q-format multiplication with adaptive quantization for tiny YOLO3. Neural Computing and Applications, 0, , .	3.2	0
291	DGL: Device Generic Latency Model for Neural Architecture Search on Mobile Devices. IEEE Transactions on Mobile Computing, 2023, , 1-14.	3.9	1
292	Deep reinforcement learning-based pairwise DNA sequence alignment method compatible with embedded edge devices. Scientific Reports, 2023, 13, .	1.6	6
293	Compression of Deep-Learning Models Through Global Weight Pruning Using Alternating Direction Method of Multipliers. International Journal of Computational Intelligence Systems, 2023, 16, .	1.6	0
294	Joint-Way Compression for LDPC Neural Decoding Algorithm With Tensor-Ring Decomposition. IEEE Access, 2023, 11, 22871-22879.	2.6	1
295	Training-aware Low Precision Quantization in Spiking Neural Networks. , 2022, , .		2

#	Article	IF	Citations
296	Neural Networks Reduction viaÂLumping. Lecture Notes in Computer Science, 2023, , 75-90.	1.0	0
297	Multi-Head Convolutional Neural Network Compression based on High-Order Principal Component Analysis. , 2023, , .		1
298	Resource-Efficient Convolutional Networks: A Survey on Model-, Arithmetic-, and Implementation-Level Techniques. ACM Computing Surveys, 2023, 55, 1-36.	16.1	2
299	Neural Network Compression for Noisy Storage Devices. Transactions on Embedded Computing Systems, 2023, 22, 1-29.	2.1	1
300	DNN Surgery: Accelerating DNN Inference on the Edge through Layer Partitioning. IEEE Transactions on Cloud Computing, 2023, , 1-15.	3.1	1
301	Neural Adaptive Loop Filtering for Video Coding: Exploring Multi-Hypothesis Sample Refinement. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33, 6057-6071.	5.6	2
302	Post-Training Quantization for Energy Efficient Realization of Deep Neural Networks. , 2022, , .		2
303	FxHENN: FPGA-based acceleration framework for homomorphic encrypted CNN inference. , 2023, , .		2
304	Mix-GEMM: An efficient HW-SW Architecture for Mixed-Precision Quantized Deep Neural Networks Inference on Edge Devices. , 2023, , .		1
305	Genie in the Model. , 2022, 7, 1-29.		0
306	Deep-learning-based multi-user framework for end-to-end fiber-MMW communications. Optics Express, 2023, 31, 15239.	1.7	3
307	Light convolutional neural network by neural architecture search and model pruning for bearing fault diagnosis and remaining useful life prediction. Scientific Reports, 2023, 13, .	1.6	6
308	Enhancing the Energy Efficiency and Robustness of tinyML Computer Vision Using Coarsely-Quantized Log-Gradient Input Images. Transactions on Embedded Computing Systems, 0, , .	2.1	2
309	PDAS: Improving network pruning based on progressive differentiable architecture search for DNNs. Future Generation Computer Systems, 2023, , .	4.9	1
310	Multi-Camera-Based Sorting System for Surface Defects of Apples. Sensors, 2023, 23, 3968.	2.1	2
311	Offloading Algorithms for Maximizing Inference Accuracy on Edge Device in an Edge Intelligence System. IEEE Transactions on Parallel and Distributed Systems, 2023, 34, 2025-2039.	4.0	0
312	Titanium oxide artificial synaptic device: Nanostructure modeling and synthesis, memristive cross-bar fabrication, and resistive switching investigation. Nano Research, 2023, 16, 10222-10233.	5.8	4
313	Global Aligned Structured Sparsity Learning for Efficient Image Super-Resolution. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45, 10974-10989.	9.7	2

#	Article	IF	Citations
314	Unsupervised Deep-Learning for Distributed Clock Synchronization in Wireless Networks. IEEE Transactions on Vehicular Technology, 2023, , 1-13.	3.9	0
317	TinyML Techniques for running Machine Learning models on Edge Devices. , 2022, , .		0
318	Compressing theÂEmbedding Matrix byÂaÂDictionary Screening Approach inÂText Classification. Lecture Notes in Computer Science, 2023, , 457-468.	1.0	0
322	In-Sensor & Neuromorphic Computing Are all You Need for Energy Efficient Computer Vision. , 2023, , .		1
324	Compiler Technologies in Deep Learning Co-Design: A Survey. , 2023, 2, .		1
325	FPGA-Based Accelerator for Rank-Enhanced and Highly-Pruned Block-Circulant Neural Networks. , 2023, , .		0
329	The Case for Hierarchical Deep Learning Inference at the Network Edge. , 2023, , .		2
331	Real-Time Unsupervised Object Localization onÂtheÂEdge forÂAirport Video Surveillance. Lecture Notes in Computer Science, 2023, , 466-478.	1.0	0
337	Hardware-Software Co-design for Side-Channel Protected Neural Network Inference. , 2023, , .		3
338	High Efficient Compression: Model Compression Method Based on Channel Pruning and Knowledge Distillation. , 2023, , .		0
341	Mixed Precision Based Parallel Optimization of Tensor Mathematical Operations on a New-generation Sunway Processor. , 2023, , .		0
343	Optimization of the Pedestrian and Vehicle Detection Model based on Cloud-Edge Collaboration. , 2022, , .		0
348	Hardware Acceleration of PPG Waveform and Heart Rate Detection System. , 2023, , .		0
362	Deep Reinforcement Learning Based Multi-Task Automated Channel Pruning for DNNs. , 2023, , .		0
363	FIANCEE: Faster Inference of Adversarial Networks via Conditional Early Exits. , 2023, , .		0
365	Architectural Vision for Quantum Computing in the Edge-Cloud Continuum. , 2023, , .		3
366	M-STREAM: A split model Streaming and Inferencing method for reduced end-to-end Execution Latency. , 2023, , .		0
371	Integer Quantized Learned Image Compression. , 2023, , .		1

#	Article	IF	Citations
374	Pruning Convolutional Filters viaÂReinforcement Learning withÂEntropy Minimization. Lecture Notes in Computer Science, 2023, , 167-180.	1.0	0
375	Slim-Tasnet: A Slimmable Neural Network for Speech Separation. , 2023, , .		0
380	Deep Learning Models Compression Based on Evolutionary Algorithms and Digital Fractional Differentiation. , 2023, , .		0
382	SignQuery: A Natural User Interface and Search Engine for Sign Languages with Wearable Sensors. , 2023, , .		1
383	Evaluating Spiking Neural Network on Neuromorphic Platform For Human Activity Recognition. , 2023, , .		0
385	A Pedestrian Detection Case Study for a Traffic Light Controller. , 2024, , 75-96.		0
386	Low Complexity OFDM-Guided DJSCC for Multipath Fading Channels using Tensor Train Decomposition with Fine-Tuning. , 2023, , .		0
396	MetaML: Automating Customizable Cross-Stage Design-Flow for Deep Learning Acceleration. , 2023, , .		1
399	Federated Boolean Neural Networks Learning. , 2023, , .		0
401	When Side-Channel Attacks Break the Black-Box Property of Embedded Artificial Intelligence. , 2023, , .		0
404	Compressing Deep Neural Networks Using Explainable AI. , 2023, , .		0
405	Review of Lightweight Deep Convolutional Neural Networks. Archives of Computational Methods in Engineering, 0, , .	6.0	0
408	Accelerating Safety Verification of Neural Network Dynamical Systems Using Assured Compressed Models. , 2023, , .		0
414	PSQ: An Automatic Search Framework for Data-Free Quantization on PIM-based Architecture. , 2023, , .		0
415	SAMP: Sub-task Aware Model Pruning withÂLayer-Wise Channel Balancing forÂPerson Search. Lecture Notes in Computer Science, 2024, , 199-211.	1.0	0
416	Accelerating Deep Neural Networks via Semi-Structured Activation Sparsity. , 2023, , .		0
417	Can Unstructured Pruning Reduce the Depth in Deep Neural Networks?. , 2023, , .		0
421	FPGA-QHAR: Throughput-Optimized for Quantized Two-Stream Human Action Recognition on the Edge. , 2023, , .		0

#	Article	IF	Citations
423	High-speed emerging memories for AI hardware accelerators. , 2024, 1, 24-34.		0
424	Reusing Deep Learning Models: Challenges and Directions in Software Engineering. , 2023, , .		0
427	A Sequence Model Compression Method Based on Proper Orthogonal Decomposition. , 2023, , .		0
428	Convolutional Neural Network Compression Based on Improved Fractal Decomposition Algorithm for Large Scale Optimization. , 2023, , .		0
431	Compressing Neural Networks with Two-Layer Decoupling. , 2023, , .		0
435	Intelligent Assisted Decision-Making Framework for Domain-Specific Advice Using Large- Language Models. , 2023, , .		0
437	Adaptive Methods for Tensor Data Compression. , 2024, , .		0
438	Resource-Limited Automated Ki67 Index Estimation in Breast Cancer. , 2023, , .		0