

Index

- 5G, 1, 2, 14, 73
- Accelerated Gradient Descent (AGD), 31
- activation functions, 14
- AdaGrad, 33
- Adaptive Moment Estimation (Adam), 35, 36
- AlexNet, 14, 15
- Ali-NPU, 96
- AllReduce, 48, 57, 138, 139
- amortized variance reduction gradient (AVRG), 32
- Apple Core ML, 173, 174
- artificial intelligence (AI), 1, 3, 78, 79, 96, 98, 107, 112, 171, 175–177, 179, 180, 182–185, 189
- Asynchronous Parallel (ASP), 22–24, 38, 151, 152, 157
- attention mechanism, 15
- autopilot, 11, 176
- backward propagation, 17, 101, 127, 171
- Batch Normalization, 37, 92
- Bert, 15, 51
- big data, 1
- big data analytics, 2–4, 8, 17, 157, 176
- Bulk Synchronous Parallel (BSP), 22, 23, 38, 82, 137, 140–143, 150–152, 156, 157
- cache servers, 12
- Central Processing Unit (CPU), 3, 6, 48, 58, 73, 78, 79, 81, 82, 96, 140, 162, 172, 174, 176, 187
- centralized learning, 1, 7, 16
- centralized server, 7, 8, 11, 51, 98, 101, 138, 140
- channel relation, 15
- CHOCO-SGD, 50
- cloud computing, 2, 11–13, 175, 183, 185
- cloud-edge environment, 3–5, 7, 128, 189
- Cloudlet, 12
- COCO, 14
- collaborative learning, 8, 163, 185, 187
- communication compression, 42, 50, 58, 63, 69, 71, 72
- communication overhead, 1, 42, 43, 45, 48, 50, 51, 53, 54, 57, 70–72, 81, 101, 103–105, 124, 130, 139, 140, 157
- communication-efficient, 5, 10, 42, 46, 53, 54, 66, 69, 70
- community based synchronization, 143–146, 149–152, 154, 156, 157
- computation acceleration, 10, 73, 79, 95
- computation capability, 16, 62, 107, 150, 177
- computation offloading, 92, 95, 108, 130, 184
- computational complexity, 11, 20, 92, 93
- computer vision, 14, 15, 189
- computing entity, 4, 5
- content distribution networks (CDN), 12
- convergence analysis, 20, 72
- convergence rate, 16, 24–27, 33, 37, 38, 46–48, 50, 53–57, 59, 60, 62, 64, 72, 101, 109, 132, 138, 140, 145–147, 149, 150, 152, 154, 156, 157
- convolutional layer, 14
- convolutional neural network (CNN), 14, 15, 17, 25, 45, 93, 118, 121
- Cyclic Repetition, 85, 87, 88
- data aggregation, 2, 134
- data center, 1, 3, 12, 94, 162, 181, 185–187
- data distribution, 6, 7, 10, 102–104, 107–109, 121, 132
- data island, 2, 8, 98, 112
- data parallelism, 10, 15, 42, 94, 131–133, 136
- data poisoning, 113, 115, 117, 118
- decentralized learning, 20, 22, 50
- decentralized parallel stochastic gradient descent (D-PSGD), 21, 138
- decision tree, 14, 133
- Deep Gradient Compression (DGC), 46
- deep learning model, 1, 10, 11, 17, 24, 25, 48, 51, 70, 93, 94, 96, 116–118, 122, 178, 183, 189
- deep neural network (DNN), 14, 15, 18, 27, 29, 37, 38, 48, 51, 57, 75, 81, 82, 92, 94, 95, 113, 117
- deep reinforcement learning, 80, 159, 162, 164, 168, 169, 187
- DeepSqueeze, 50
- difference compression SGD (DCD), 50
- differential privacy (DP), 122–125
- Directed Acyclic Graph, 139
- dist-EF-SGD, 50

- distributed machine learning, 1–3, 15, 23, 24, 33, 62, 73, 90, 112, 120, 124, 157
- distributed momentum SGD (DMSGD), 47
- distributed SGD, 24–27, 47, 53, 57–59, 71, 72, 150
- edge computing, 3, 4, 11, 12, 94, 161, 175, 179, 181, 182, 185, 189
- edge intelligence, 4, 131, 133, 137, 163, 171–175
- edge learning architecture, 20
- Efficient Multiple Instance learning (EMI-RNN), 93
- error feedback, 48
- Error-Compensated SGD, 49
- Euclidean distance, 28, 29
- European Telecommunications Standards Institute (ETSI), 14
- extrapolation compression SGD (ECD), 50
- fast analog transmission (FAT), 68
- feature maps, 15
- federated learning, 3, 4, 8, 16, 47, 51, 52, 62, 63, 68, 70, 98–101, 103, 104, 106–111, 163, 172, 174, 175, 183
- Field Programmable Gate Array (FPGA), 73, 78, 79, 81, 82, 94, 96
- finer-grained aggregation, 52
- fog computing, 13
- forward propagation, 77
- fully homomorphic encryption, 126
- fundamental theory, 10, 109
- game theory, 160, 161, 165
- Generative Adversarial Networks (GAN), 37, 104, 118, 121, 122, 127
- GhostNet, 15
- global momentum compression (GMC), 47
- GoogleNet, 15
- gradient aggregation, 45, 109, 110, 138, 144
- Gradient Centralization, 37
- gradient coding, 82–84, 90, 95, 97, 110
- Gradient Descent (GD), 17, 25, 30, 164
- gradient quantization, 42, 43, 53, 57, 58, 70, 71
- GradiVeQ, 45
- Graphic Processing Unit (GPU), 51, 58, 73, 74, 76–82, 94, 96, 140, 172, 174, 187
- Graphical Models, 3
- Hessian matrix, 72
- heterogeneous devices, 6, 8, 16
- hit rate, 12
- homomorphic encryption (HE), 125–127, 130
- Hyper-Adam, 37
- hyper-parameter, 19, 24, 27, 29, 33, 34, 36, 37, 49, 53, 100, 107, 109, 110, 137
- ImageNet, 3, 14, 15, 93
- incentive mechanism, 6, 8, 10, 159–165, 168–170
- independent and identically distributed (IID), 7, 101–104, 107, 108, 132, 133, 165
- Internet of Things (IoT), 1, 2, 11, 12, 69, 132, 137, 172, 173, 175, 177, 179–181, 184, 185
- K-AVG SGD, 54
- K-nearest neighbors, 14
- Layer Normalization (LN), 37
- lazy synchronization, 51
- linear discriminant analysis (LDA), 46
- linear minimum mean square error (LMMSE), 69
- linear regression, 17, 24, 25, 90, 113, 118, 121, 124, 126, 133
- Lipschitzness, 37
- local data, 8, 100–102, 104, 112, 126, 133, 138, 164, 168, 175
- local update, 6
- logistic regression, 14, 17, 24, 25, 108, 121, 124, 133
- long and short-term memory (LSTM), 15, 93
- low rank factorization, 43, 74, 75, 92, 95
- machine learning algorithm, 11, 46, 73, 107, 108
- machine learning application, 8, 19, 74–80, 108, 131, 133, 139, 157, 173, 177, 179
- machine learning model, 1, 3–5, 14, 15, 17, 24, 42, 43, 47, 51, 70, 73, 90, 93, 96, 100, 101, 113, 128, 140, 161, 163–165, 168, 172, 173, 183, 189
- machine learning system, 8, 74, 79, 80, 82, 84–86, 96, 131, 136–138
- macro base station (MBS), 139, 141–144, 150
- malicious attacker, 71, 128, 130
- malicious attacks, 8
- matrix factorization, 75
- Mean Squared Error (MSE), 28
- Message Passing Interface (MPI), 80
- micro-clouds, 12
- Mini-batch SGD, 19, 29, 34, 52, 142
- mini-batch SGD, 17
- mobile computing, 11, 12, 180
- mobile edge computing (MEC), 13, 14, 108, 130, 140
- MobileNet, 15
- model accuracy, 10, 78, 91, 96, 117, 123, 125, 127, 163, 165
- model parallelism, 10, 15, 94, 131, 133, 135, 136
- model pruning, 117
- Momentum SGD, 24
- multi-instance learning (MIL), 93
- multiple-input multiple-output (MIMO), 66
- multiuser multiple-input and multiple-output (MU-MIMO), 68

- natural language processing, 14, 15, 42
- Nesterov-accelerated Adaptive Momentum Estimation (Nadam), 35
- network congestion, 11, 12
- network resources, 12
- network traffic, 11, 77, 138
- Neural Network, 3
- Neural Processing Unit (NPU), 73, 82, 95, 96, 180
- Non-IID, 7, 16, 101–105, 107, 108
- non-IID, 101
- Non-Orthogonal Multiple Access (NOMA), 66
- NVIDIA Collective Communications Library (NCCL), 80, 172
- NVIDIA Deep Learning Accelerator (NVDLA), 96
- one-shot aggregation, 52
- OpenFog, 13
- outlier detection, 130
- over-fitting, 117, 133, 175
- Overlap Synchronization Parallel (OSP), 23
- parallel computing, 8, 157
- parallel restarted, 53
- parameter server, 16, 20–24, 38, 46, 50, 52, 54, 55, 57, 58, 63, 65, 66, 68, 69, 98, 107, 110, 136–138, 140, 151, 156, 158, 164–170
- peer-to-peer (P2P), 12, 16
- periodic averaging, 51–53
- Physical-Layer Arithmetic (PhyArith), 68
- Polynomial Coded Regression, 90
- pooling layer, 14
- PowerSGD, 48
- practical application, 1, 70, 143, 171
- principal component analysis (PCA), 45, 123
- privacy leakage, 7, 174, 177, 188, 189
- privacy-sensitive, 3, 4, 7
- PyTorch, 171, 172
- PyTorch Mobile, 172
- QSGD, 43–45, 50, 54, 57, 61, 156
- quantization, 43
- Quanzited Overlap Synchronization Parallel (QOSP), 42, 58–62
- radio access network (RAN), 14
- radio frequency (RF), 69
- raw data, 5, 7, 90, 99, 101, 107, 108, 121, 125
- receptive fields (RFs), 15
- recommendation system, 1, 14, 162
- recurrent neural networks (RNN), 15, 37, 93, 94, 109, 110, 172, 178
- regularization, 27–29, 36, 37, 48, 116, 118
- reinforcement learning, 81, 82, 108–110, 115
- ResNet, 77
- resource constraint, 5, 6
- resource utilization, 3, 10, 73, 94, 174
- resource-intensive, 12, 81
- response delay, 1, 2, 11, 185
- risk loss, 27, 28
- SAGA, 31
- SARAH, 32
- SARAH++, 32
- self-driving, 1, 162, 171
- SENet, 15
- sequential D-SGD, 62
- service quality, 11–13, 180, 185
- SignSGD, 43
- Singular Value Decomposition (SVD), 75
- SKNet, 15
- small base station (SBS), 139–144, 150
- smart grid, 1, 161, 175
- smart healthcare, 1, 171, 175, 179, 180, 188
- smart home, 1
- smart surveillance, 1, 171
- smart transportation, 171, 175–177, 179
- social network, 4
- sparification, 43, 46–48, 50, 57, 63, 71, 72, 76, 156
- SpiderBoost, 32
- SqueezeNet, 15
- Stackelberg equilibrium, 118, 166–169
- Stackelberg game, 118, 161, 162, 164, 165, 168, 169
- Stale Synchronous Parallel (SSP), 22–24, 38, 136
- staleness, 23, 24, 38, 46, 60, 62, 135, 151, 152, 154
- Stochastic Average Gradient (SAG), 31
- Stochastic Gradient Descent (SGD), 17–20, 24, 25, 29–31, 34, 36, 38, 42, 46, 52, 64, 70–72, 100, 101, 109, 125, 139, 144, 146, 147, 156
- Stochastic Variance Reduced Gradient (SVRG), 31
- straggler tolerance, 73
- Sufficient Factor Broadcasting (SFB), 75
- SUNRGBD, 14
- support vector machine (SVM), 14, 28, 101, 125, 133
- synchronization mode, 10, 11, 23
- Tensor Processing Unit (TPU), 81, 82, 96
- TensorFlow, 172, 173
- TensorFlow Lite, 173
- TernGrad, 43
- training algorithm, 10, 16, 24, 118, 122, 126, 127
- unbiased estimation, 18, 20
- VGG-16, 3
- Virtual Reality/Augmented Reality, 11
- Webank FATE, 175
- Weight Normalization (WN), 37
- XLNet, 15
- zeroth-order Adam method (ZO-AdaMM), 35