

Index

- accountability, 166
- accuracy, 140
- actions, 18, 19
 - model, contrasted, 18
 - predictions, contrasted, 18
- activation function, 171, 172
- activations, 171
- actuator, 19
- AdaGrad, 296
- Adam, 296
- adaptive overfitting, 152
- additive attention scoring, 430
- affine transformation, 83
- agent, 18
- AI, 26, 27
 - general, 27
- ALBERT, 465
- AlexNet, 273, 275, 283, 286
- AlphaGo, 26
- ALU (arithmetic logical unit), 275
- AnyNet, 321
- argmax, 140
- armed bandit problem, 20
- array, 30
- artificial general intelligence, 27
- attention, 416
 - multi-head, 437
- attention mechanism, 23, 414, 416
 - Bahdanau, 432
- attention pooling, 417, 420, 433
- attention scoring, 425
 - additive, 430
 - distance-based, 431
 - scaled dot product, 426, 428
- attention weight, 417
- attribute, *see* feature
- AUC (area under the receiving operating characteristic), 55
- audio data, 242
- autograd, *see* automatic differentiation
- automatic differentiation, 60, 63, 64
- autoregression, 379, 447
- autoregressive model, 331, 332
 - latent, 332
- average, *see* mean
- average pooling, 260, 264
- axon, 91
- axon terminal, 91
- backpropagation, 22, 59, 60, 183, 184
 - sequence model, 366
- backpropagation through time, 366, 369
 - truncated, 368
- backward differentiation, 65
- backward propagation, *see* backpropagation
- bags of visual words, 272
- Bahdanau attention mechanism, 432
- bandit problem
 - armed, 20
 - contextual, 20
 - multi-armed, 20
- BART, 465
- batch learning, 164
- batch matrix multiplication (BMM), 428
- batch normalization, 295, 296, 298
- Bayes' theorem, 71
- Bayesian probability, 65
- beam search, 410, 412
- beam size, 412
- Bernoulli random variable, 150
- BERT, 463
 - fine-tuning, 464
 - pretraining, 463
- BFLOAT16, 148

- bias, 83
 bidirectional encoder representations from
 transformers, *see* BERT
 bidirectional RNN, 390, 403
 implementation, 392
 BIG-Bench, 471
 bigram, 347
 bilingual evaluation understudy, *see* BLEU
 binary classification, 10
 BLEU, 408, 435
 BLOOM, 472
 Boxcar kernel, 420, 422
 break the symmetry, 188
 broadcasting, 35
 C4 (Colossal Clean Crawled Corpus), 466
 Caffe, 25, 108
 Caffe2, 25
 calculus, 54
 candidate hidden state, 382
 category, 10
 causal attention, 465
 causality, 17
 censored feedback, 14
 central limit theorem, 68, 150
 chain rule, 59, 183
 chain-of-thought prompting, 472
 Chainer, 25
 channel, 240
 dimension, 254
 input, 255
 output, 256
 character-level language model, 355
 ChatGPT, 24, 472
 Chebyshev's inequality, 77
 Chinchilla, 471
 CIFAR, 274
 class, *see* category
 classification, 10, 126, 127
 accuracy, 139
 binary, 10
 hierarchical, 11
 multi-label, 12, 127
 multiclass, 10
 CLIP (contrastive language-image pretrain-
 ing), 473
 clustering, 16
 CNN, 23, 235, 239, 328
 CNTK, 25
 co-adaptation, 196
 column vectors, 42
 comma-separated values (CSV), 38
 complexity of a class, 192
 complexity of a function, 119
 component of a vector, *see* element
 computational graph, 60, 182, 184
 concept shift, 158
 correction, 163
 conditional probability, 71
 conditionally independent random variables,
 73
 confusion matrix, 163
 consistency, 66
 constant parameter, 214
 context variable, 402, 433
 contextual bandit problem, 20
 control flow, 63
 control theory, 165
 convolution, 238, 239, 247
 1×1 , 257, 286
 grouped, 312
 strided, 250
 convolution window, *see* kernel window
 convolutional kernel, 239
 convolutional layer, 239
 convolutional neural network, *see* CNN
 covariance, 76
 covariance matrix, 76
 covariate, *see* feature
 covariate shift, 156
 correction, 161
 internal, 303
 CPU, 229
 credit assignment, 19
 cross-correlation, 240, 247
 cross-entropy, 10, 132
 cross-entropy loss, 130, 131, 133, 142
 cross-validation, K -fold, 117, 118, 205
 CUDA, 487
 d2l package, xxvii
 DALL-E 2, 24, 473

- data
 - dimensionality, 5
 - fixed-length, 5
 - varying-length, 5
- data point, *see* example
- data transfer, 231
- data example, 5
- dataset, 3
 - size, 116
 - test, 7
 - training, 7, 82
 - validation, 87, 105
- decoder, 399
- deep generative model, 17
- deep generative modeling, 24
- deep learning, 2, 6, 22, 27, 116, 148
 - framework, 25
- deep neural network, 270
- deep Q-network, 19
- deep reinforcement learning, 19
- deep RNN, 386
 - implementation, 389
- DeIT, 462
- deletion, 39
- dendrite, 91
- dense block, 317
- DenseNet, 315
- derivative, 54
- design matrix, 84
- detaching, 62
- differentiability, 55
- differentiation
 - backward, 65
 - forward, 65
- diffusion model, 24
- dimensionality of a tensor, 43
- dimensionality of a vector, 42
- dimensionality of data, 5
- DistBelief, 108
- DistilBERT, 465
- Distributed Shampoo, 296
- distribution
 - nonstationary, 159
- distribution of a random variable, 70
- distribution shift, 18, 156
 - correction, 160
- dot product, 48
- double-descent pattern, 192
- dropout, 23, 196, 449
 - implementation, 197
 - probability, 196
- early stopping, 193, 195
- EfficientNet, 321
- ELECTRA, 465
- elements of a vector, 42
- embedding layer, 403
- empirical error, 149, 153
- empirical risk, 160
- empirical risk minimization, 161
 - weighted, 161
- encoder, 399
- encoder–decoder architecture, 399
- end-to-end training, 28
- energy-based models, 129
- entropy, 132
- entry of a vector, *see* element
- Epanechnikov kernel, 420, 422
- epoch, 105, 108
- error
 - empirical, 149, 153
 - generalization, 113–115, 191, 192
 - out-of-memory, 184
 - population, 150
 - training, 113, 114
 - true, 153
 - validation, 115, 123
- estimation
 - mean, 114, 150
- estimator, 66
 - maximum likelihood, 90
- event, 69
- example, 5, 8, 82
 - category, 10
- exhaustive search, 410, 411
- expectation of a random variable, 74
- exploding gradient, 186, 187, 366
- exploitation of strategies, 20
- exploration of strategies, 20
- fairness, 166

- FALCON, 472
 Fashion-MNIST, 135–137, 200
 feature, 5, 65, 82
 dimensionality, 90
 feature engineering, 28
 feature map, 240, 247
 feature selection, 120
 file I/O, 228
 filter, 239
 fitness, 84
 Flamingo, 473
 forget gate, 375
 forward differentiation, 65
 forward pass, *see* forward propagation
 forward propagation, 181, 184
 foundation model, 281, 415
 FP16, 148
 FP32 single precision, 148
 FP64 double precision, 148
 frequency of an event, 67
 frequentism, 65
 Galactica, 471
 games, 26
 gate
 forget, 375
 input, 375
 output, 375
 reset, 381
 update, 381
 gated recurrent unit, *see* GRU
 gating, 374
 Gato, 463
 Gaussian distribution, *see* normal distribution
 Gaussian error linear units, *see* GELU
 Gaussian kernel, 420, 422, 426
 GELU, 176, 459
 generalization, 54, 87, 113, 118, 193, 194
 generalization error, 113–115, 191, 192
 generalization gap, 114, 115, 191
 generative adversarial network (GAN), 17, 24
 generative pretraining, *see* GPT
 Gluon, 25
 GoogLeNet, 289, 291, 294, 309, 320
 Gopher, 471
 GPT, 471
 GPT-2, 468
 GPT-3, 468, 469
 GPT-4, 470
 GPU, 228, 229, 274
 tensor storage, 230
 gradient, 58, 131
 clipping, 361
 exploding, 186, 187, 366
 vanishing, 186, 187, 366
 gradient descent, 7, 86
 stochastic, 22, 25, 86, 107
 graphical processing unit, *see* GPU
 greedy search, 407, 410
 ground truth, 7, 8, 52, 100, 107, 140
 grouped convolution, 312
 growth rate, 318
 GRU, 373, 381
 implementation, 384
 Hadamard product, \odot , 45
 Hebbian learning rule, 22
 hidden layer, 168, 170
 hidden representation, 170
 hidden state, 352, 353
 candidate, 382
 GRU, 383
 LSTM, 376
 hierarchical classification, 11
 Hoeffding's inequality, 151
 HOG, 272
 hyperbolic tangent, *see* tanh
 hyperparameter, 87, 106, 143
 optimization, 321
 hypothesis testing
 multiple, 152
 identity mapping, 307
 IID, 113, 118
 Imagen, 24
 ImageNet, 135, 274
 imperative tools, 25
 imputation, 39
 in place updating, 36
 in-context learning, 469

- few-shot, 469
- one-shot, 469
- zero-shot, 469
- Inception, 312
- Inception block, 290
- independent random variables, 72
- inductive bias, 191
- inference, 87
- information theory, 21, 131
- inner product, *see* dot product
- input, 5
 - channel, 255
 - gate, 375
 - node, 375
- InstructGPT, 472
- INT8, 148
- integer-location based indexing (iloc), 39
- intercept, *see* bias
- intermediate value, 183
- internal covariate shift, 303
- intra-attention model, 441
- iteration, 105
- Jacobian, 62
- JAX, 25, 111
- joint probability, 71
- Jupyter Notebook, xxviii, 94, 474, 488
- Kaggle, 200
- Keras, 25
- kernel density estimation, 420
- kernel methods, 193
- kernel window, 242
- key, 416, 428
- Kryder's law, 22
- label, 5, 82
- label shift, 157
 - correction, 162
- LAION, 274
- LAION-400m, 327
- LAION-5B, 327
- language model, 24, 333, 346
 - character-level, 355
 - large, 471
 - masked, 463
- Laplace smoothing, 347
- lasso regression, 120
- lasso regularization, 194
- law of large numbers, 68
- layer, 22, 27
 - convolutional, 239
 - embedding, 403
 - fully-connected, 109, 128
 - hidden, 168, 170
 - lazy, 109
 - normalization, 298
 - transition, 317
- layer normalization, 448
- lazy initialization, 221
- learning
 - algorithm, 3
 - batch, 164
 - deep, 22, 27
 - in-context, 469
 - offline, 17
 - online, 164
 - rate, 86
 - reinforcement, 19, 25, 165
 - deep, 19
 - self-supervised, 17
 - sequence, 15
 - supervised, 4, 8, 65, 126
 - unsupervised, 16
- LeNet, 265, 271, 275
 - implementation, 266
- likelihood, 72, 90
 - maximum, 90
- linear model, 169
- linear regression, 83
 - high dimension, 121
- linear transformation, 83
- Linux, xxvi
- Lipschitz continuous, 360
- Llama 1, 472
- Llama 2, 472
- locality, 237, 238
- log–log plot, 343
- log-likelihood
 - negative, 90, 130
- LogSumExp trick, 147
- long short-term memory, *see* LSTM

- loss function, 6, 84
 squared error, 7, 84
 LSTM, 23, 374
 machine learning, 1, 2, 6, 21, 65, 107, 151
 accountability, 166
 fairness, 166
 transparency, 166
 machine translation, 15, 401
 neural, 393
 statistical, 393
 macOS, xxvi
 MAE, 465
 Manhattan distance, *see* norm, ℓ_1
 marginalization, 72
 Markov decision process, 20
 Markov model, 333
 k^{th} -order, 333
 masked language model, 463
 masked softmax, 427
 masking, 406
 matplotlib library, 56
 matrix, 31, 43
 invertible, 85
 square, 43
 symmetric, 44
 transpose, 44
 matrix multiplication, 49, 50
 matrix–matrix multiplication, *see* matrix multiplication
 matrix–vector product, 49
 max-pooling, 264
 maximum likelihood
 principle of, 90
 estimator, 90
 maximum pooling, *see* max-pooling
 mean, 47
 mean estimation, 21
 Megatron–Turing NLG, 471
 memory cell, 374, 376
 output, 376
 memory network, 24
 Minerva, 472
 minibatch, 86
 minibatch stochastic gradient descent, 120, 188
 MLP, 168, 170, 171, 176, 191, 210, 352
 hidden layer, 170
 implementation, 177
 input layer, 170
 output layer, 170
 MNIST, 135
 model, 3, 6
 expert, 462
 expressivity, 192
 family, 3
 foundation, 281
 generalist, 462
 input, 3
 output, 3
 selection, 117
 statistical, 6
 module
 class, 210
 parallel, 215
 monotonicity, 169
 Monte Carlo tree sampling, 26
 Moore’s law, 22
 multitmodal model, 470
 multi-armed bandit problem, 20
 multi-head attention, 437
 multi-label classification, 12, 127
 multiclass classification, 10
 multilayer perceptron, 23, *see* MLP
 mutually exclusive events, 69
 MXNet, 25, 111
 n -gram, 408
 Nadaraya–Watson estimator, 420
 Nadaraya–Watson regression, 422, 424
 NaN (Not a Number), 39, 146
 NAS (neural architecture search), 321
 nat, 132
 ndarray, 30
 Neocognitron, 241
 network engineering, 320
 network in network, *see* NiN
 neural network, 22, 193
 convolutional, *see* CNN
 deep, 270
 module, 210
 parametrization symmetry, 187

- recurrent, 328
- time-delay, 238
- training, 184
- neural programmer-interpreter, 24
- neuron, 91
- NiN, 285, 286, 320
- no free lunch theorem, 191
- nonparametric model, 193
- nonstationary distribution, 159
- norm, 50
 - ℓ_1 norm, 51, 119
 - ℓ_2 norm (Euclidean), 51, 119
 - ℓ_p norm, 51
 - Frobenius norm, 52
 - spectral, 51
- normal distribution, 89
- normalization, 129
 - batch, 295, 296, 298
 - layer, 298
- nucleus, 91
- numerical stability, 147, 185
- object recognition, 26
- object-oriented programming, 94
- objective function, 4, 6
 - surrogate, 7
- observation, 19
- offline learning, 17
- offset, *see* bias
- one-hot encoding, 127, 358
- online learning, 164
- Open Pretrained Transformers (OPT), 472
- operation
 - binary scalar, 34
 - elementwise, 34
 - unary scalar, 34
- operator
 - differentiation, 56
- optimization, 7
- order of a tensor, 43
- out-of-memory error, 184
- outcome, 69
- outcome space, *see* sample space
- outer product, 76
- output
 - channel, 256
 - gate, 375
- overfitting, 7, 23, 113, 115, 119, 122, 191
 - adaptive, 152
- overflow, 146
- padding, 250, 262, 396
- PaLM, 471
- PaLM 2, 471
- pandas, 38
- parallel module, 215
- parameter, 3, 23
 - constant, 214
 - initialization, 188
 - model, 84
 - scale, 297
 - shift, 297
 - tied, 218
- parametrized ReLU, *see* pReLU
- ParNet, 284
- Parti, 473
- partial derivative, 58
- partition function
 - log-, 134, 141
- Parzen window, 420, 424
- Pathway Language Model, *see* PaLM
- patience criterion, 194
- pattern, 113
 - double-descent, 192
- perceptron, 22
 - convergence theorem, 295
 - multilayer, 23, 168
- perplexity, 349
- personalization, 13
- point-of-speech (POS) tagging, 15
- Poisson distribution, 93
- policy, 19
- polysemy, 392
- pooling, 260
 - average pooling, 260
 - layer, 260
 - max-pooling, 260
 - window, 260
- population error, 150
- positional encoding, 442
- post-normalization, 460
- posterior belief, 72

- power-law scaling, 470
- pre-normalization, 460
- prediction, 87, 107, 336
 - k -step ahead, 337
- prediction mode, 297
- pReLU, 173, 176
- pretraining, 462
- principal components analysis, 17
- prior, 72
- probabilistic graphical models, 17
- probability, 65
 - conditional, 71
 - function, 69
 - joint, 71
- probit model, 129
- prod operator, 183
- prompting, 472
 - chain-of-thought, 472
- proportional-integral-derivative (PID) control, 165
- PyTorch, xxvii, 25, 30, 111
- Q-learning, 23
- query, 416, 428
- radial basis function (RBF), 171
- random variable, 69
 - Bernoulli, 150
 - conditionally independence, 73
 - continuous, 70
 - discrete, 70
 - distribution, 70
 - expectation, 74
 - independence, 72
 - variance, 75
- receptive field, 247
- recommender system, 13
- rectified linear unit, *see* ReLU
- recurrent layer, 353
- recurrent neural network, *see* RNN
- reduction of a tensor, 46
- RegNet, 321
- RegNetX, 326
- RegNetY, 326
- regression, 9, 82, 126
 - linear, 83
- regularization, 113, 119, 125, 194, 306
 - ℓ_2 , *see* weight decay
- reinforcement learning, 19, 25, 165
 - deep, 19
 - from AI feedback, 472
 - from human feedback, 472
- ReLU, 171, 172, 187, 264, 277
- reset gate, 381
- residual block, 307
- residual connection, 307, 448
- residual mapping, 307
- residual network, *see* ResNet
- ResNet, 306, 309, 320
- ResNeXt, 312, 320
- ResNeXt block, 312
- ridge regression, 120
- ridge regularization, 194
- risk, 160
 - empirical, 160
- RKHS (reproducing kernel Hilbert space), 124, 171
- RNN, 328, 352, 374
 - bidirectional, 390, 403
 - decoder, 404
 - deep, 386
 - encoder, 403
 - encoder–decoder, 405
 - attention, 433
 - hidden state, 353
- RoBERTa, 465
- runaway feedback loop, 166
- sample, *see* example
- sample space, 69
- sampling, 67
- scalability, 24, 25, 462, 470
- scalar, 41
- scale parameter, 297
- scaled dot product attention scoring, 426, 428
- scaling
 - power-law, 470
- search
 - beam, 410, 412
 - exhaustive, 410, 411
 - greedy, 407, 410

- self-attention model, 441
- self-driving vehicles, 26
- self-supervised learning, 17
- SENet, 320, 419
- seq2seq, *see* sequence-to-sequence
- sequence, 330
 - learning, 15
 - model, 331, 333
 - source, 393
 - target, 393
- sequence-to-sequence, 393
 - aligned, 331
 - learning, 401, 453
 - encoder–decoder, 406
 - unaligned, 331
- shape, 31, 45
 - of a tensor, 31, 34
- shift parameter, 297
- ShiftNet, 314
- shortcut connection, *see* residual connection
- SIFT, 28, 271
- sigmoid function, 173, 174, 177, 186, 277
 - derivative, 174
- slice, 36
- slope of a function, 56
- SN2, 108
- softmax, 128, 129, 131, 133, 134, 141
 - masked, 427
- source distribution, 161
- SpanBERT, 465
- speech recognition, 15, 26
- squared error, 7
- squashing function, 173
- Squeeze-and-Excitation Network, *see* SENet
- SSH port forwarding, 488
- standard deviation, 75
- standardization, 203
- stationary dynamics, 333
- statistical learning theory, 152
- statistical physics, 129
- statistics, 66
- stochastic gradient descent (SGD), 86, 107
 - minibatch, 86
- stop word, 343
- stride, 252, 262
- strided convolution, 250
- subspace estimation, 17
- supervised learning, 4, 8, 65, 126
- SURF, 271
- surprisal, 132
- surrogate objective, 7
- survival modeling, 126
- Swin Transformer, 462, 465
- Swish activation function, 176
- synapses, 91
- T5, 465
 - fine-tuning, 466
 - pretraining, 465
- tagging, 11
 - part-of-speech (POS), 15
- tanh, 175, 177
 - derivative, 175
- target, 5, 65, 82
- target distribution, 161
- teacher forcing, 402
- tensor, 30, 44
 - k^{th} -order, 31
 - arithmetic, 45
 - element, 31
 - reduction, 46
 - shape, 31
- TensorFlow, 25, 111
- test data, 18
- test set, 7
- Text-to-Text Transfer Transformer, *see* T5
- Theano, 25, 108
- theory of computation, 21
- tied parameters, 218
- Tikhonov regularization, 195
- time-delay neural network (TDNN), 238
- TinyImages, 274
- token, 333, 340
- tokenization
 - character-level, 340
 - word-level, 395
- Torch, 25
- TP32, 148
- training
 - data, 18
 - dataset, 82

- error, 113, 114
- loop, 87, 105
- mode, 297
- set, *see* training dataset
- Transformer, 24, 320, 327, 414, 415, 446
 - decoder, 451
 - decoder only, 467
 - encoder, 449
 - encoder only, 463
 - encoder–decoder, 453, 465
 - implementation, 449
 - pretraining, 462
 - scalability, 470
 - vision, *see* ViT
- transition layer, 317
- translation, 83
- translation equivariance, *see* translation invariance
- translation invariance, 237, 238
- transparency, 166
- transpose of a matrix, 44
- trigram, 347
- trimmed mean estimate, 21
- trimming, 267
- true error, 153
- truncation, 368, 396
 - randomized, 368
- uncertainty, 65
 - aleatoric, 76
 - epistemic, 76
- underfitting, 115
- underflow, 146
- unigram, 347
- unsupervised density modeling, 331
- unsupervised learning, 16
- update gate, 381
- utility, 75
- validation
 - dataset, 87, 105, 117
 - error, 115, 123
 - set, *see* validation dataset
- value, 416, 428
- vanishing gradient, 186, 366
- variance of a random variable, 75
- variational autoencoders, 17
- VC (Vapnik–Chervonenkis) dimension, 153
- vector, 30, 42
 - column, 42
 - elements, 42
 - row, 42
- vectorization, 130
- VGG, 281–283, 286, 307, 320
- ViT (vision Transformer), 457
 - encoder, 459
 - implementation
 - patch embedding, 459
 - training, 461
- vocabulary, 341
- warm-up period, 362
- web search, 13
- weight, 83
 - synaptic, 91
- weight decay, 115, 119, 121, 123, 194
- Windows, xxvi
- WordNet, 274
- Xavier initialization, 188, 189
- Yogi, 296
- zero-based indexing, 42
- Zipf’s law, 343