Index

1D convolution, 66

2D convolution, 66

activation, 23, 41

    function, 71, 99

    map, 69

Adam, 39, 154

adapter, 153

affine operation, 61

artificial neural network, 8, 11

attention operator, 88

autoencoder, 159

    denoising, 118

Autograd, 42

autoregressive model, *see* model, autoregressive

average pooling, 76

backpropagation, 42

backward pass, 42, 154

basis function regression, 14

batch, 21, 38

batch normalization, 80, 104

Bellman equation, 135
bias vector, 61, 67
BPE, *see* Byte Pair Encoding
Byte Pair Encoding, 34, 129

cache memory, 21
capacity, 16
causal, 32, 90, 111
    model, *see* model, causal
chain rule (derivative), 40
chain rule (probability), 30
chain-of-thought, 140, 149
channel, 23
checkpointing, 43
classification, 18, 26, 101, 120
CLIP, *see* Contrastive Language-Image Pre-training
CLS token, 115
computational cost, 43, 91
context size, 147
Contrastive Language-Image Pre-training, 131, 156
contrastive loss, 27, 131
convnet, *see* convolutional network
convolution, 66
convolutional layer, *see* layer, convolutional
convolutional network, 101
cross-attention block, 94, 109, 111
cross-entropy, 27, 31, 45

data augmentation, 120
deep learning, 8, 11
Deep Q-Network, 135
denoising autoencoder, *see* autoencoder, denoising
density modeling, 18
depth, 41
diffusion model, 142
dilation, 67, 74
discriminator, 160
downscaling residual block, 106
downstream task, 50
DQN, *see* Deep Q-Network
dropout, 77, 91

embedding layer, *see* layer, embedding
epoch, 48
equivariance, 67, 94

feed-forward block, 108, 109
few-shot prediction, 139
filter, 66
fine-tune, 124
fine-tuning, 51, 141
flops, 22
forward pass, 41
foundation model, 140
FP32, 22
framework, 23

GAN, *see* Generative Adversarial Networks
GELU, 73
Generative Adversarial Networks, 160
Generative Pre-trained Transformer, 112, 131, 139, 162
generator, 160
GNN, *see* Graph Neural Network
GPT, *see* Generative Pre-trained Transformer
GPU, *see* Graphical Processing Unit
gradient descent, 35, 37, 40, 45
gradient norm clipping, 44
gradient step, 35
Graph Neural Network, 161
Graphical Processing Unit, 8, 20
ground truth, 18

hidden layer, *see* layer, hidden
hidden state, 158
hyper parameter, *see* parameter, hyper
hyperbolic tangent, 72

image processing, 101
image synthesis, 87, 142
inductive bias, 17, 49, 66, 67, 96
invariance, 76, 94, 96, 162

kernel size, 66, 74
key, 88

Large Language Model, 51, 56, 88, 139, 146, 162

layer, 41, 59
- attention, 87
- convolutional, 66, 74, 87, 96, 101, 104, 121, 126, 129
- embedding, 95, 111
- fully connected, 61, 87, 96, 99, 101
- hidden, 99
- linear, 61
- Multi-Head Attention, 93, 96, 111
- normalizing, 80
- reversible, 43

layer normalization, 83, 108, 111
Leaky ReLU, 72
learning rate, 35, 50
learning rate schedule, 50
LeNet, 101, 102
linear layer, *see* layer, linear
LLM, *see* Large Language Model
local minimum, 35
logit, 26, 31
LoRA, *see* Low-Rank Adaptation
loss, 12
Low-Rank Adaptation, 153, 155

machine learning, 11, 17, 18
Markovian Decision Process, 134
Markovian property, 134
max pooling, 74, 101
MDP, *see* Markovian, Decision Process

mean squared error, 14, 26
memory requirement, 43
memory speed, 21
metric learning, 27
MLP, *see* multi-layer perceptron, 154
model, 12
    autoregressive, 30, 31, 139
    causal, 33, 91, 111, 112
    parametric, 12
    pre-trained, 51, 124, 128
model merging, 156
multi-layer perceptron, 45, 99–101, 108

Natural Language Processing, 87
NLP, *see* Natural Language Processing
non-linearity, 71
normalizing layer, *see* layer, normalizing

object detection, 121
overfitting, 17, 48

padding, 67, 74
parameter, 12
    hyper, 13, 35, 48, 66, 67, 74, 93, 95
parametric model, *see* model, parametric
peak performance, 22
Perplexity, 151
perplexity, 31
policy, 134
    optimal, 134

pooling, 74
positional encoding, 96, 111
Post-Training Quantization, 150
posterior probability, 26
pre-trained model, *see* model, pre-trained
prompt, 139, 140
    engineering, 147

quantization, 150
Quantization-Aware Training, 152
query, 88

RAG, *see* Retrieval-Augmented Generation
random initialization, 62
receptive field, 68, 69, 124
rectified linear unit, 71, 158
recurrent neural network, 158
regression, 18
Reinforcement Learning, 134, 141
Reinforcement Learning from Human Feedback, 141
ReLU, *see* rectified linear unit
residual
    block, 104
    connection, 84, 103
    network, 47, 84, 103
ResNet-50, 103
Retrieval-Augmented Generation, 149
return, 134

reversible layer, *see* layer, reversible
RL, *see* Reinforcement Learning
RLHF, *see* Reinforcement Learning from Human Feeback
RNN, *see* recurrent neural network

scaling laws, 52
self-attention block, 94, 109, 111
self-supervised learning, 162
semantic segmentation, 86, 126
SGD, *see* stochastic gradient descent
Single Shot Detector, 121
skip connection, 84, 127, 158
softargmax, 26, 89
softmax, 26
speech recognition, 129
SSD, *see* Single Shot Detector
stochastic gradient descent, 38, 45, 52
stride, 67, 74
supervised learning, 19

Tanh, *see* hyperbolic tangent
Task Arithmetic, 156
tensor, 23
tensor cores, 21
Tensor Processing Unit, 21
test set, 48
text synthesis, 139
token, 30

tokenizer, 34, 129
TPU, *see* Tensor Processing Unit
trainable parameter, 12, 23, 52
training, 12
training set, 12, 25, 48
Transformer, 47, 84, 88, 96, 108, 110, 129
transformer, 154
transposed convolution, 69, 126

underfitting, 16
universal approximation theorem, 99
unsupervised learning, 19

VAE, *see* variational, autoencoder
validation set, 48
value, 88
vanishing gradient, 44, 58
variational
    autoencoder, 160
    bound, 144
Vision Transformer, 113, 131
ViT, *see* Vision Transformer
vocabulary, 30

weight, 13
    decay, 28
    matrix, 61

zero-shot prediction, 132

This book is licensed under the Creative Commons BY-NC-SA 4.0 International License.

V1.2–May 19, 2024