

Contents

1	Foundations	1
1.1	Neural networks	1
1.1.1	Overview	1
1.1.2	Components	2
1.2	Training	5
1.2.1	Parameter learning	5
1.2.2	Optimizers	11
1.2.3	Common loss functions	16
1.2.4	Regularization	20
1.3	Evaluation	23
1.3.1	Data splits	23
1.3.2	Metrics	25
1.3.3	Bias-variance trade-off	29
2	Embeddings	33
2.1	Tokenization	33
2.1.1	Notations	33
2.1.2	Subword algorithms	36
2.2	Token embeddings	40
2.2.1	One-hot encodings	40
2.2.2	Continuous encodings	41
2.3	Document embeddings	46
2.3.1	Heuristic methods	46
2.3.2	Recurrent neural networks	49
2.3.3	Attention-based methods	57
2.4	Embedding operations	59
2.4.1	Similarity	59
2.4.2	Dimension reduction	60
2.4.3	Fast retrieval	64
3	Transformers	69
3.1	Self-attention mechanism	69
3.1.1	Query, key, value	69
3.1.2	Multi-head attention	71
3.1.3	Transformer architecture	73
3.1.4	End-to-end example	76
3.2	Components	83
3.2.1	Input	84

3.2.2	Encoder	86
3.2.3	Decoder	88
3.2.4	Output	90
3.3	Computational improvements	92
3.3.1	Position awareness	92
3.3.2	Layer-related tricks	98
3.4	Architecture variations	102
3.4.1	BERT and encoder-only models	102
3.4.2	GPT and decoder-only models	109
3.4.3	T5 and encoder-decoder models	112
3.4.4	Switch Transformer and mixture of expert models	114
3.5	Attention computations speedup	118
3.5.1	Sparse attention	118
3.5.2	Low rank attention	123
3.5.3	Hardware optimization	125
3.6	Interpretability	128
3.6.1	Attention-specific methods	128
3.6.2	Model-agnostic methods	129
4	Large language models	139
4.1	Fundamentals	139
4.1.1	Notations	139
4.1.2	Response generation	141
4.2	Pretraining	146
4.2.1	Data mixtures	146
4.2.2	Training objective	147
4.3	Prompt engineering	147
4.3.1	Basics	147
4.3.2	Strategies	149
4.3.3	Risks	155
4.4	Finetuning	156
4.4.1	Supervised finetuning	157
4.4.2	Parameter efficient finetuning	159
4.5	Preference tuning	164
4.5.1	Concepts	165
4.5.2	Reinforcement learning from human feedback	167
4.5.3	Supervised preference tuning	171
4.6	Model compression	173
4.6.1	Overview	173
4.6.2	Distillation	176
4.6.3	Quantization	177

5	Applications	189
5.1	Evaluation framework	189
5.1.1	General concepts	191
5.1.2	Common benchmarks	195
5.2	Classification tasks	195
5.2.1	Sentiment extraction	196
5.2.2	Named entity recognition	197
5.3	Generation tasks	197
5.3.1	Machine translation	204
5.3.2	Summarization	208
5.3.3	Retrieval-augmented generation	208
	Index	225