
Contents

Foreword		xvii
Preface		xix
Authors		xxiii
Contributors		xxv
CHAPTER 1 ■ Deep Learning and Transformers: An Introduction		1
1.1	DEEP LEARNING: A HISTORIC PERSPECTIVE	1
1.2	TRANSFORMERS AND TAXONOMY	4
1.2.1	Modified Transformer Architecture	4
1.2.1.1	Transformer block changes	4
1.2.1.2	Transformer sublayer changes	5
1.2.2	Pre-training Methods and Applications	8
1.3	RESOURCES	8
1.3.1	Libraries and Implementations	8
1.3.2	Books	9
1.3.3	Courses, Tutorials, and Lectures	9
1.3.4	Case Studies and Details	10
CHAPTER 2 ■ Transformers: Basics and Introduction		11
2.1	ENCODER-DECODER ARCHITECTURE	11
2.2	SEQUENCE-TO-SEQUENCE	12
2.2.1	Encoder	12

2.2.2	Decoder	13
2.2.3	Training	14
2.2.4	Issues with RNN-Based Encoder-Decoder	14
2.3	ATTENTION MECHANISM	14
2.3.1	Background	14
2.3.2	Types of Score-Based Attention	16
2.3.2.1	Dot product (multiplicative)	17
2.3.2.2	Scaled dot product or multiplicative	17
2.3.2.3	Linear, MLP, or Additive	17
2.3.3	Attention-Based Sequence-to-Sequence	18
2.4	TRANSFORMER	19
2.4.1	Source and Target Representation	20
2.4.1.1	Word embedding	20
2.4.1.2	Positional encoding	20
2.4.2	Attention Layers	22
2.4.2.1	Self-attention	22
2.4.2.2	Multi-head attention	24
2.4.2.3	Masked multi-head attention	25
2.4.2.4	Encoder-decoder multi-head attention	26
2.4.3	Residuals and Layer Normalization	26
2.4.4	Positionwise Feed-forward Networks	26
2.4.5	Encoder	27
2.4.6	Decoder	27
2.5	CASE STUDY: MACHINE TRANSLATION	27
2.5.1	Goal	27
2.5.2	Data, Tools, and Libraries	27
2.5.3	Experiments, Results, and Analysis	28
2.5.3.1	Exploratory data analysis	28
2.5.3.2	Attention	29
2.5.3.3	Transformer	35
2.5.3.4	Results and analysis	38
2.5.3.5	Explainability	38

CHAPTER	3 ■ Bidirectional Encoder Representations from Transformers (BERT)	43
3.1	BERT	43
3.1.1	Architecture	43
3.1.2	Pre-Training	45
3.1.3	Fine-Tuning	46
3.2	BERT VARIANTS	48
3.2.1	RoBERTa	48
3.3	APPLICATIONS	49
3.3.1	TaBERT	49
3.3.2	BERTopic	50
3.4	BERT INSIGHTS	51
3.4.1	BERT Sentence Representation	51
3.4.2	BERTology	52
3.5	CASE STUDY: TOPIC MODELING WITH TRANSFORMERS	53
3.5.1	Goal	53
3.5.2	Data, Tools, and Libraries	53
3.5.2.1	Data	54
3.5.2.2	Compute embeddings	54
3.5.3	Experiments, Results, and Analysis	55
3.5.3.1	Building topics	55
3.5.3.2	Topic size distribution	55
3.5.3.3	Visualization of topics	56
3.5.3.4	Content of topics	57
3.6	CASE STUDY: FINE-TUNING BERT	63
3.6.1	Goal	63
3.6.2	Data, Tools, and Libraries	63
3.6.3	Experiments, Results, and Analysis	64

CHAPTER	4 ■ Multilingual Transformer Architectures	71
4.1	MULTILINGUAL TRANSFORMER ARCHITECTURES	72
4.1.1	Basic Multilingual Transformer	72
4.1.2	Single-Encoder Multilingual NLU	74
4.1.2.1	mBERT	74
4.1.2.2	XLM	75
4.1.2.3	XLM-RoBERTa	77
4.1.2.4	ALM	77
4.1.2.5	Unicoder	78
4.1.2.6	INFOXML	80
4.1.2.7	AMBER	81
4.1.2.8	ERNIE-M	82
4.1.2.9	HITCL	84
4.1.3	Dual-Encoder Multilingual NLU	85
4.1.3.1	LaBSE	85
4.1.3.2	mUSE	87
4.1.4	Multilingual NLG	89
4.2	MULTILINGUAL DATA	90
4.2.1	Pre-Training Data	90
4.2.2	Multilingual Benchmarks	91
4.2.2.1	Classification	91
4.2.2.2	Structure prediction	92
4.2.2.3	Question answering	92
4.2.2.4	Semantic retrieval	92
4.3	MULTILINGUAL TRANSFER LEARNING INSIGHTS	93
4.3.1	Zero-Shot Cross-Lingual Learning	93
4.3.1.1	Data factors	93
4.3.1.2	Model architecture factors	94
4.3.1.3	Model tasks factors	95
4.3.2	Language-Agnostic Cross-Lingual Representations	96

4.4	CASE STUDY	97
4.4.1	Goal	97
4.4.2	Data, Tools, and Libraries	98
4.4.3	Experiments, Results, and Analysis	98
4.4.3.1	Data preprocessing	99
4.4.3.2	Experiments	101
CHAPTER	5 ■ Transformer Modifications	109
<hr/>		
5.1	TRANSFORMER BLOCK MODIFICATIONS	109
5.1.1	Lightweight Transformers	109
5.1.1.1	Funnel-transformer	109
5.1.1.2	DeLight	112
5.1.2	Connections between Transformer Blocks	114
5.1.2.1	RealFormer	114
5.1.3	Adaptive Computation Time	115
5.1.3.1	Universal transformers (UT)	115
5.1.4	Recurrence Relations between Transformer Blocks	116
5.1.4.1	Transformer-XL	116
5.1.5	Hierarchical Transformers	120
5.2	TRANSFORMERS WITH MODIFIED MULTI-HEAD SELF-ATTENTION	120
5.2.1	Structure of Multi-Head Self-Attention	120
5.2.1.1	Multi-head self-attention	122
5.2.1.2	Space and time complexity	123
5.2.2	Reducing Complexity of Self-Attention	124
5.2.2.1	Longformer	124
5.2.2.2	Reformer	126
5.2.2.3	Performer	131
5.2.2.4	Big Bird	132
5.2.3	Improving Multi-Head-Attention	137
5.2.3.1	Talking-heads attention	137
5.2.4	Biasing Attention with Priors	140

5.2.5	Prototype Queries	140
5.2.5.1	Clustered attention	140
5.2.6	Compressed Key-Value Memory	141
5.2.6.1	Luna: Linear Unified Nested Attention	141
5.2.7	Low-Rank Approximations	143
5.2.7.1	Linformer	143
5.3	MODIFICATIONS FOR TRAINING TASK EFFICIENCY	145
5.3.1	ELECTRA	145
5.3.1.1	Replaced token detection	145
5.3.2	T5	146
5.4	TRANSFORMER SUBMODULE CHANGES	146
5.4.1	Switch Transformer	146
5.5	CASE STUDY: SENTIMENT ANALYSIS	148
5.5.1	Goal	148
5.5.2	Data, Tools, and Libraries	148
5.5.3	Experiments, Results, and Analysis	150
5.5.3.1	Visualizing attention head weights	150
5.5.3.2	Analysis	152
CHAPTER 6	Pre-trained and Application-Specific Transformers	155
6.1	TEXT PROCESSING	155
6.1.1	Domain-Specific Transformers	155
6.1.1.1	BioBERT	155
6.1.1.2	SciBERT	156
6.1.1.3	FinBERT	156
6.1.2	Text-to-Text Transformers	157
6.1.2.1	ByT5	157
6.1.3	Text Generation	158
6.1.3.1	GPT: Generative pre-training	158
6.1.3.2	GPT-2	160
6.1.3.3	GPT-3	161

6.2	COMPUTER VISION	163
6.2.1	Vision Transformer	163
6.3	AUTOMATIC SPEECH RECOGNITION	164
6.3.1	Wav2vec 2.0	165
6.3.2	Speech2Text2	165
6.3.3	HuBERT: Hidden Units BERT	166
6.4	MULTIMODAL AND MULTITASKING TRANSFORMER	166
6.4.1	Vision-and-Language BERT (ViLBERT)	167
6.4.2	Unified Transformer (UniT)	168
6.5	VIDEO PROCESSING WITH TIMESFORMER	169
6.5.1	Patch Embeddings	169
6.5.2	Self-Attention	170
6.5.2.1	Spatiotemporal self-attention	171
6.5.2.2	Spatiotemporal attention blocks	171
6.6	GRAPH TRANSFORMERS	172
6.6.1	Positional Encodings in a Graph	173
6.6.1.1	Laplacian positional encodings	173
6.6.2	Graph Transformer Input	173
6.6.2.1	Graphs without edge attributes	174
6.6.2.2	Graphs with edge attributes	175
6.7	REINFORCEMENT LEARNING	177
6.7.1	Decision Transformer	178
6.8	CASE STUDY: AUTOMATIC SPEECH RECOGNITION	180
6.8.1	Goal	180
6.8.2	Data, Tools, and Libraries	180
6.8.3	Experiments, Results, and Analysis	180
6.8.3.1	Preprocessing speech data	180
6.8.3.2	Evaluation	181

CHAPTER 7	Interpretability and Explainability Techniques for Transformers	187
7.1	TRAITS OF EXPLAINABLE SYSTEMS	187
7.2	RELATED AREAS THAT IMPACT EXPLAINABILITY	189
7.3	EXPLAINABLE METHODS TAXONOMY	190
7.3.1	Visualization Methods	190
7.3.1.1	Backpropagation-based	190
7.3.1.2	Perturbation-based	194
7.3.2	Model Distillation	195
7.3.2.1	Local approximation	195
7.3.2.2	Model translation	198
7.3.3	Intrinsic Methods	198
7.3.3.1	Probing mechanism	198
7.3.3.2	Joint training	201
7.4	ATTENTION AND EXPLANATION	202
7.4.1	Attention is Not an Explanation	202
7.4.1.1	Attention weights and feature importance	202
7.4.1.2	Counterfactual experiments	204
7.4.2	Attention is Not Not an Explanation	205
7.4.2.1	Is attention necessary for all tasks?	206
7.4.2.2	Searching for adversarial models	207
7.4.2.3	Attention probing	208
7.5	QUANTIFYING ATTENTION FLOW	208
7.5.1	Information Flow as DAG	208
7.5.2	Attention Rollout	209
7.5.3	Attention Flow	209
7.6	CASE STUDY: TEXT CLASSIFICATION WITH EXPLAINABILITY	210
7.6.1	Goal	210
7.6.2	Data, Tools, and Libraries	211
7.6.3	Experiments, Results, and Analysis	211

