

Table of Contents

Foreword.....	xi
Preface.....	xv
1. Hello Transformers.....	1
The Encoder-Decoder Framework	2
Attention Mechanisms	4
Transfer Learning in NLP	6
Hugging Face Transformers: Bridging the Gap	9
A Tour of Transformer Applications	10
Text Classification	10
Named Entity Recognition	11
Question Answering	12
Summarization	13
Translation	13
Text Generation	14
The Hugging Face Ecosystem	15
The Hugging Face Hub	16
Hugging Face Tokenizers	17
Hugging Face Datasets	18
Hugging Face Accelerate	18
Main Challenges with Transformers	19
Conclusion	20
2. Text Classification.....	21
The Dataset	22
A First Look at Hugging Face Datasets	23
From Datasets to DataFrames	26

Looking at the Class Distribution	27
How Long Are Our Tweets?	28
From Text to Tokens	29
Character Tokenization	29
Word Tokenization	31
Subword Tokenization	33
Tokenizing the Whole Dataset	35
Training a Text Classifier	36
Transformers as Feature Extractors	38
Fine-Tuning Transformers	45
Conclusion	54
3. Transformer Anatomy.....	57
The Transformer Architecture	57
The Encoder	60
Self-Attention	61
The Feed-Forward Layer	70
Adding Layer Normalization	71
Positional Embeddings	73
Adding a Classification Head	75
The Decoder	76
Meet the Transformers	78
The Transformer Tree of Life	78
The Encoder Branch	79
The Decoder Branch	82
The Encoder-Decoder Branch	83
Conclusion	85
4. Multilingual Named Entity Recognition.....	87
The Dataset	88
Multilingual Transformers	92
A Closer Look at Tokenization	93
The Tokenizer Pipeline	94
The SentencePiece Tokenizer	95
Transformers for Named Entity Recognition	96
The Anatomy of the Transformers Model Class	98
Bodies and Heads	98
Creating a Custom Model for Token Classification	99
Loading a Custom Model	101
Tokenizing Texts for NER	103
Performance Measures	105
Fine-Tuning XLM-RoBERTa	106

Error Analysis	108
Cross-Lingual Transfer	115
When Does Zero-Shot Transfer Make Sense?	116
Fine-Tuning on Multiple Languages at Once	118
Interacting with Model Widgets	121
Conclusion	122
5. Text Generation.....	123
The Challenge with Generating Coherent Text	125
Greedy Search Decoding	127
Beam Search Decoding	130
Sampling Methods	134
Top-k and Nucleus Sampling	136
Which Decoding Method Is Best?	140
Conclusion	140
6. Summarization.....	141
The CNN/DailyMail Dataset	141
Text Summarization Pipelines	143
Summarization Baseline	143
GPT-2	144
T5	144
BART	145
PEGASUS	145
Comparing Different Summaries	146
Measuring the Quality of Generated Text	148
BLEU	148
ROUGE	152
Evaluating PEGASUS on the CNN/DailyMail Dataset	154
Training a Summarization Model	157
Evaluating PEGASUS on SAMSum	158
Fine-Tuning PEGASUS	158
Generating Dialogue Summaries	162
Conclusion	163
7. Question Answering.....	165
Building a Review-Based QA System	166
The Dataset	167
Extracting Answers from Text	173
Using Haystack to Build a QA Pipeline	181
Improving Our QA Pipeline	189
Evaluating the Retriever	189

Evaluating the Reader	196
Domain Adaptation	199
Evaluating the Whole QA Pipeline	203
Going Beyond Extractive QA	205
Conclusion	207
8. Making Transformers Efficient in Production.....	209
Intent Detection as a Case Study	210
Creating a Performance Benchmark	212
Making Models Smaller via Knowledge Distillation	217
Knowledge Distillation for Fine-Tuning	217
Knowledge Distillation for Pretraining	220
Creating a Knowledge Distillation Trainer	220
Choosing a Good Student Initialization	222
Finding Good Hyperparameters with Optuna	226
Benchmarking Our Distilled Model	229
Making Models Faster with Quantization	230
Benchmarking Our Quantized Model	236
Optimizing Inference with ONNX and the ONNX Runtime	237
Making Models Sparser with Weight Pruning	243
Sparsity in Deep Neural Networks	244
Weight Pruning Methods	244
Conclusion	248
9. Dealing with Few to No Labels.....	249
Building a GitHub Issues Tagger	251
Getting the Data	252
Preparing the Data	253
Creating Training Sets	257
Creating Training Slices	259
Implementing a Naive Bayesline	260
Working with No Labeled Data	263
Working with a Few Labels	271
Data Augmentation	271
Using Embeddings as a Lookup Table	275
Fine-Tuning a Vanilla Transformer	284
In-Context and Few-Shot Learning with Prompts	288
Leveraging Unlabeled Data	289
Fine-Tuning a Language Model	289
Fine-Tuning a Classifier	293
Advanced Methods	295
Conclusion	297

10. Training Transformers from Scratch.....	299
Large Datasets and Where to Find Them	300
Challenges of Building a Large-Scale Corpus	300
Building a Custom Code Dataset	303
Working with Large Datasets	306
Adding Datasets to the Hugging Face Hub	309
Building a Tokenizer	310
The Tokenizer Model	312
Measuring Tokenizer Performance	312
A Tokenizer for Python	313
Training a Tokenizer	318
Saving a Custom Tokenizer on the Hub	322
Training a Model from Scratch	323
A Tale of Pretraining Objectives	323
Initializing the Model	325
Implementing the Dataloader	326
Defining the Training Loop	330
The Training Run	337
Results and Analysis	338
Conclusion	343
11. Future Directions.....	345
Scaling Transformers	345
Scaling Laws	347
Challenges with Scaling	349
Attention Please!	351
Sparse Attention	352
Linearized Attention	353
Going Beyond Text	354
Vision	355
Tables	359
Multimodal Transformers	361
Speech-to-Text	361
Vision and Text	364
Where to from Here?	370
Index.....	371