

УДК 004.4  
ББК 32.972  
И29

**Саваш Йылдырым, Мейсам Асгари-Ченаглу**

**И29** Осваиваем архитектуру Transformer. Разработка современных моделей с помощью передовых методов обработки естественного языка / пер. с англ. В. С. Яценкова. – М.: ДМК Пресс, 2022. – 320 с.: ил.

**ISBN 978-5-93700-106-1**

Основанные на трансформерах языковые модели – преобладающая тема исследований в области обработки естественного языка (NLP). В этой книге рассказывается, как создавать различные приложения NLP на основе трансформеров, используя библиотеку Python Transformers.

Вы познакомитесь с архитектурой трансформеров и напишете свою первую программу для работы с моделями на основе этой передовой технологии.

Книга адресована специалистам по NLP, преподавателям машинного обучения / NLP и тем, кто хочет освоить машинное обучение в части обработки естественного языка. Предполагается, что читатель владеет навыками программирования на языке Python, знает основы NLP и понимает, как работают глубокие нейронные сети.

УДК 004.4  
ББК 32.972

Copyright ©Packt Publishing 2021. First published in the English language under the title 'Mastering Transformers - (9781801077651)

Все права защищены. Любая часть этой книги не может быть воспроизведена в какой бы то ни было форме и какими бы то ни было средствами без письменного разрешения владельцев авторских прав.

ISBN (англ.) 978-1-80107-765-1  
ISBN (рус.) 978-5-93700-106-1

© 2021 Packt Publishing  
© Оформление, издание, перевод, ДМК Пресс, 2022

# Оглавление

<b>Об авторах .....</b>	<b>5</b>
<b>О рецензенте .....</b>	<b>6</b>
<b>Предисловие .....</b>	<b>11</b>
Для кого эта книга .....	11
Какие темы охватывает эта книга .....	11
Как получить максимальную отдачу от этой книги .....	12
Скачивание исходного кода примеров .....	13
Видеоролики Code in Action .....	13
Условные обозначения и соглашения, принятые в книге .....	13
Список опечаток .....	14
Нарушение авторских прав .....	14
 <b>ЧАСТЬ I. ПОСЛЕДНИЕ РАЗРАБОТКИ В ОБЛАСТИ NLP, ПОДГОТОВКА РАБОЧЕЙ СРЕДЫ И ПРИЛОЖЕНИЕ HELLO WORLD .....</b>	 <b>15</b>
<b>Глава 1. От последовательности слов к трансформерам .....</b>	<b>17</b>
Технические требования .....	18
Эволюция подходов NLP в направлении трансформеров .....	18
Что такое дистрибутивная семантика? .....	21
Использование глубокого обучения .....	26
Обзор архитектуры трансформеров .....	37
Трансформеры и перенос обучения .....	46
Заключение .....	48
Дополнительная литература .....	48
<b>Глава 2. Знакомство с трансформерами на практике .....</b>	<b>49</b>
Технические требования .....	50
Установка библиотеки Transformer с Anaconda .....	51
Работа с языковыми моделями и токенизаторами .....	57

Работа с моделями, предоставленными сообществом .....	59
Сравнительное тестирование и наборы данных .....	62
Тестирование быстродействия и использования памяти .....	74
Заключение .....	77

## ЧАСТЬ II. МОДЕЛИ-ТРАНСФОРМЕРЫ – ОТ АВТОЭНКОДЕРОВ К АВТОРЕГРЕССИИ.....79

### Глава 3. Языковые модели на основе автоэнкодеров.....81

Технические требования.....	82
BERT – одна из языковых моделей на основе автоэнкодера .....	82
Обучение автоэнкодерной языковой модели для любого языка.....	86
Как поделиться моделями с сообществом .....	97
Обзор других моделей с автоэнкодером.....	98
Использование алгоритмов токенизации .....	104
Заключение .....	115

### Глава 4. Авторегрессивные и другие языковые модели.....116

Технические требования.....	117
Работа с языковыми моделями AR.....	117
Работа с моделями Seq2Seq.....	122
Обучение авторегрессивной языковой модели .....	127
Генерация текста с использованием авторегрессивных моделей.....	132
Тонкая настройка резюмирования и машинного перевода с помощью simpletransformers .....	135
Заключение .....	138
Дополнительная литература.....	138

### Глава 5. Тонкая настройка языковых моделей для классификации текста.....139

Технические требования.....	140
Введение в классификацию текста.....	140
Тонкая настройка модели BERT для двоичной классификации с одним предложением .....	141
Обучение модели классификации с помощью PyTorch.....	148
Тонкая настройка BERT для многоклассовой классификации с пользовательскими наборами данных.....	152
Тонкая настройка BERT для регрессии пар предложений.....	158

Использование <code>run_glue.py</code> для тонкой настройки моделей .....	163
Заключение .....	164
<b>Глава 6. Тонкая настройка языковых моделей для классификации токенов.....</b>	<b>165</b>
Технические требования .....	166
Введение в классификацию токенов.....	166
Тонкая настройка языковых моделей для NER .....	171
Ответы на вопросы с использованием классификации токенов .....	179
Заключение .....	187
<b>Глава 7. Представление текста .....</b>	<b>188</b>
Технические требования .....	188
Введение в представление предложений .....	189
Эксперимент по выявлению семантического сходства с FLAIR .....	198
Кластеризация текста с помощью Sentence-BERT .....	204
Семантический поиск с помощью Sentence-BERT .....	209
Заключение .....	213
Дополнительная литература.....	214
<b>ЧАСТЬ III. ДОПОЛНИТЕЛЬНЫЕ ТЕМЫ .....</b>	<b>215</b>
<b>Глава 8. Работа с эффективными трансформерами .....</b>	<b>217</b>
Технические требования .....	218
Обзор эффективных, легких и быстрых трансформеров .....	218
Способы уменьшения размера модели.....	220
Работа с эффективным самовниманием .....	226
Заключение .....	246
Дополнительная литература.....	247
<b>Глава 9. Многоязычные и кросс-языковые модели .....</b>	<b>248</b>
Технические требования .....	249
Моделирование языка перевода и обмен знаниями между языками.....	249
XLM и mBERT .....	251
Задачи выявления кросс-языкового сходства .....	256
Кросс-языковая классификация.....	263
Кросс-языковое обучение без подготовки .....	268
Фундаментальные ограничения многоязычных моделей .....	271
Заключение .....	274
Дополнительная литература.....	274

## Глава 10. Трансформерная модель

<b>как самостоятельная служба</b> .....	275
Технические требования.....	276
Запуск службы трансформерной модели с fastAPI.....	276
Докеризация API.....	279
Создание службы модели с использованием TFX.....	280
Нагрузочное тестирование службы с помощью Locust.....	282
Заключение .....	286
Дополнительные источники информации .....	286

## Глава 11. Визуализация внимания и отслеживание

<b>экспериментов</b> .....	287
Технические требования.....	288
Интерпретация механизма внимания.....	288
Многоуровневая визуализация потоков внимания с помощью BertViz.....	294
Заключение .....	312
Дополнительная литература.....	313

<b>Предметный указатель</b> .....	314
-----------------------------------	-----