Оглавление

Предисловие ... 11
Благодарности ... 13
О книге ... 15
Кому следует прочитать эту книгу ... 15
Структура издания ... 16
О коде ... 17
Другие онлайн-ресурсы ... 18
Об авторе ... 19
Иллюстрация на обложке ... 20
От издательства ... 21
О научном редакторе русскоязычного издания ... 21

Глава 1. Знакомство с большими языковыми моделями ... 22
1.1. Что такое LLM ... 23
1.2. Применение больших языковых моделей ... 25
1.3. Зачем создавать собственные модели ... 27
1.4. Этапы обучения LLM ... 27
1.5. Введение в архитектуру трансформера ... 29
1.6. Использование больших выборок данных ... 33
1.7. Более подробный обзор архитектуры GPT ... 34
1.8. План создания LLM с нуля ... 36
Итоги главы ... 38

Глава 2. Работа с текстовыми данными ... 40
2.1. Векторные представления слов ... 42
2.2. Токенизация текста ... 45
2.3. Преобразование токенов в идентификаторы токенов ... 49
2.4. Добавление специальных контекстных токенов ... 53
2.5. Кодирование пар байтов ... 57
2.6. Выборка данных с помощью скользящего окна ... 60
2.7. Создание векторных представлений токенов ... 66
2.8. Кодирование позиций слов ... 70
Итоги главы ... 75

Глава 3. Программирование механизмов внимания ... 76
3.1. Проблема моделирования длинных последовательностей ... 78
3.2. Улавливание зависимостей данных с помощью механизмов внимания ... 81
3.3. Обращение к разным частям входных данных с помощью самовнимания ... 83
3.3.1. Простой механизм самовнимания без обучаемых весов ... 83
3.3.2. Вычисление весов внимания для всех входных токенов ... 89
3.4. Реализация самовнимания с обучаемыми весами ... 91
3.4.1. Пошаговое вычисление весовых коэффициентов внимания ... 93
3.4.2. Реализация компактного класса Python для самовнимания ... 98
3.5. Сокрытие будущих слов с помощью причинно-следственного внимания ... 102
3.5.1. Применение маски причинно-следственного внимания ... 103
3.5.2. Маскирование дополнительных весов внимания с помощью отсева ... 107
3.5.3. Реализация компактного класса причинно-следственного внимания ... 109
3.6. Расширение одноцелевого внимания до многоцелевого ... 111
3.6.1. Объединение нескольких одноцелевых слоев внимания ... 112
3.6.2. Реализация многоцелевого внимания с разделением весов ... 115
Итоги главы ... 121

Глава 4. Создание GPT-подобной модели для генерации текста с нуля 122
4.1. Программирование архитектуры LLM 124
4.2. Нормализация активаций с помощью нормализации слоев 130
4.3. Реализация сети с прямым распространением и активацией GELU 136
4.4. Добавление коротких соединений 140
4.5. Объединение механизма внимания и линейных слоев в блоке трансформера 144
4.6. Программирование модели GPT 148
4.7. Генерация текста 154
Итоги главы 159

Глава 5. Предварительное обучение на неразмеченных данных 160
5.1. Оценка генеративных текстовых моделей 162
5.1.1. Использование GPT для генерации текста 163
5.1.2. Расчет потерь при генерации текста 165
5.1.3. Расчет потерь для обучающей и проверочной выборок 173
5.2. Обучение LLM 179
5.3. Стратегии декодирования для контроля случайности 185
5.3.1. Масштабирование температуры 186
5.3.2. Выборка по принципу top-k 190
5.3.3. Изменение функции генерации текста 192
5.4. Загрузка и сохранение весов модели в PyTorch 193
5.5. Загрузка предварительно обученных весов из OpenAI 195
Итоги главы 203

Глава 6. Тонкая настройка по классификации 204
6.1. Различные категории тонкой настройки 206
6.2. Подготовка данных 208
6.3. Создание загрузчиков данных 211
6.4. Инициализация модели с предварительно обученными весами 218
6.5. Добавление цели классификации 220
6.6. Расчет потерь и точности классификации 228
6.7. Тонкая настройка модели на данных с метками 232
6.8. Использование LLM в качестве классификатора спама 239
Итоги главы 241

Глава 7. Тонкая настройка по инструкциям ... 242
7.1. Введение в тонкую настройку по инструкциям ... 244
7.2. Подготовка набора данных для контролируемой тонкой настройки по инструкциям ... 244
7.3. Организация данных в обучающие пакеты ... 249
7.4. Создание загрузчиков данных для набора инструкций ... 263
7.5. Загрузка предварительно обученной LLM ... 266
7.6. Тонкая настройка LLM по инструкциям ... 269
7.7. Извлечение и сохранение ответов ... 274
7.8. Оценка точно настроенной LLM ... 279
7.9. Выводы ... 288
7.9.1. Что дальше ... 290
7.9.2. Будьте в курсе последних событий в быстро меняющейся области ... 290
7.9.3. Заключительные слова ... 290
Итоги главы ... 291

Приложение A. Введение в PyTorch ... 292
A.1. Что такое PyTorch ... 292
A.1.1. Три основных компонента PyTorch ... 293
A.1.2. Определение глубокого обучения ... 294
A.1.3. Установка PyTorch ... 296
A.2. Тензоры ... 300
A.2.1. Скаляры, векторы, матрицы и тензоры ... 301
A.2.2. Типы тензорных данных ... 301
A.2.3. Распространенные операции с тензорами в PyTorch ... 302
A.3. Представление моделей в виде графов вычислений ... 304
A.4. Автоматическое дифференцирование упростилось ... 305
A.5. Реализация многослойных нейронных сетей ... 308
A.6. Настройка эффективных загрузчиков данных ... 313
A.7. Типичный цикл обучения ... 318
A.8. Сохранение и загрузка моделей ... 323
A.9. Оптимизация производительности обучения с помощью графических процессоров ... 323
A.9.1. Вычисления PyTorch на графических процессорах ... 324

A.9.2. Обучение с использованием одного графического процессора ... 325
A.9.3. Обучение с использованием нескольких графических процессоров ... 327
Итоги приложения ... 334

Приложение Б. Ссылки и дополнительные источники ... 335

Приложение В. Решения упражнений ... 347

Приложение Г. Добавление новых возможностей в процесс обучения ... 359
Г.1. Переменная скорость обучения ... 360
Г.2. Затухание по косинусу ... 362
Г.3. Ограничение градиента ... 364
Г.4. Измененная функция обучения ... 365

Приложение Д. Эффективная настройка параметров с помощью LoRA ... 368
Д.1. Введение в LoRA ... 368
Д.2. Подготовка набора данных ... 370
Д.3. Инициализация модели ... 373
Д.4. Эффективная тонкая настройка параметров с помощью LoRA ... 375