# Оглавление


Часть I ■ ОСНОВЫ МАШИННОГО ОБУЧЕНИЯ ..... 25
*1* ■ Конструирование современного машинного обучения ..... 26
*2* ■ Глубокие нейронные сети ..... 46
*3* ■ Сверточная и остаточная нейронные сети ..... 75
*4* ■ Основы процесса тренировки ..... 106

Часть II ■ БАЗОВЫЙ ШАБЛОН КОНСТРУИРОВАНИЯ ..... 163
*5* ■ Шаблон процедурного конструирования ..... 165
*6* ■ Широкие сверточные нейронные сети ..... 199
*7* ■ Альтернативные шаблоны связности ..... 235
*8* ■ Мобильные сверточные нейронные сети ..... 263
*9* ■ Автокодировщики ..... 309

Часть III ■ РАБОТА С КОНВЕЙЕРАМИ ..... 336
*10* ■ Гиперпараметрическая настройка ..... 338
*11* ■ Перенос обучения ..... 369
*12* ■ Распределения данных ..... 396
*13* ■ Конвейер данных ..... 420
*14* ■ Конвейер тренировки и развертывания ..... 467

# Содержание

*Предисловие* ... 13
*Признательности* ... 14
*Об этой книге* ... 15
*Об авторе* ... 22
*Об иллюстрации на обложке* ... 24

**Часть I ОСНОВЫ МАШИННОГО ОБУЧЕНИЯ** ... 25

1 ***Конструирование современного машинного обучения*** ... 26
1.1 Курс на адаптируемость ... 27
*1.1.1 Компьютерное зрение задает тон* ... 29
*1.1.2 За пределами компьютерного зрения: обработка ЕЯ, понимание ЕЯ, структурированные данные* ... 30
1.2 Эволюция подходов, основанных на машинном обучении ... 31
*1.2.1 Классический ИИ против узкого ИИ* ... 31
*1.2.2 Следующие шаги в компьютерном обучении* ... 35
1.3 Выгоды от шаблонов конструирования ... 42
Резюме ... 45

2 ***Глубокие нейронные сети*** ... 46
2.1 Основы нейронных сетей ... 47
*2.1.1 Входной слой* ... 47
*2.1.2 Глубокие нейронные сети* ... 50
*2.1.3 Сети прямого распространения* ... 51
*2.1.4 Метод последовательного API* ... 51
*2.1.5 Метод функционального API* ... 52
*2.1.6 Входная форма и входной слой* ... 52
*2.1.7 Плотный слой* ... 53
*2.1.8 Активационные функции* ... 55
*2.1.9 Сокращенный синтаксис* ... 59

*2.1.10 Повышение точности с помощью оптимизатора* ... 60
2.2 Двоичный классификатор в форме глубокой нейронной сети ... 61
2.3 Мультиклассовый классификатор в форме глубокой нейронной сети ... 63
2.4 Мультиметочный мультиклассовый классификатор в форме глубокой нейронной сети ... 66
2.5 Простой классификатор изображений ... 68
*2.5.1 Разглаживание* ... 69
*2.5.2 Переподгонка и отсев* ... 71
Резюме ... 73

**3 *Сверточная и остаточная нейронные сети*** ... 75
3.1 Сверточные нейронные сети ... 76
*3.1.1 Зачем для моделирования изображений использовать сверточную нейросеть поверх глубокой нейросети* ... 77
*3.1.2 Отбор с пониженной частотой (изменение размера)* ... 77
*3.1.3 Обнаружение признаков* ... 79
*3.1.4 Сведение* ... 82
*3.1.5 Разглаживание* ... 83
3.2 Конструкция в форме ConvNet для сверточной нейросети ... 83
3.3 Сети в форме VGG ... 88
3.4 Сети в форме ResNet ... 92
*3.4.1 Архитектура* ... 93
*3.4.2 Пакетная нормализация* ... 99
*3.4.3 Архитектура ResNet50* ... 100
Резюме ... 104

**4 *Основы процесса тренировки*** ... 106
4.1 Прямая подача и обратное распространение ... 107
*4.1.1 Подача данных* ... 108
*4.1.2 Обратное распространение* ... 108
4.2 Разбивка набора данных ... 110
*4.2.1 Тренировочный и тестовый наборы* ... 111
*4.2.2 Кодирование с одним активным состоянием* ... 113
4.3 Нормализация данных ... 116
*4.3.1 Нормализация* ... 116
*4.3.2 Стандартизация* ... 118
4.4 Валидация и переподгонка ... 119
*4.4.1 Валидация* ... 119
*4.4.2 Слежение за потерей* ... 123
*4.4.3 Погружение вглубь с помощью слоев* ... 123
4.5 Схождение ... 125
4.6 Фиксация контрольных точек и ранняя остановка ... 128
*4.6.1 Фиксация контрольной точки* ... 128
*4.6.2 Ранняя остановка* ... 130
4.7 Гиперпараметры ... 131
*4.7.1 Эпохи* ... 132

*4.7.2 Шаги* ..... 132
*4.7.3 Размер пакета* ..... 134
*4.7.4 Скорость усвоения* ..... 135
4.8 Инвариантность ..... 138
*4.8.1 Трансляционная инвариантность* ..... 140
*4.8.2 Масштабная инвариантность* ..... 147
*4.8.3 ImageDataGenerator модуля TF.Keras* ..... 148
4.9 Сырые (дисковые) наборы данных ..... 150
*4.9.1 Каталожная структура* ..... 151
*4.9.2 Файл CSV* ..... 153
*4.9.3 Файл JSON* ..... 154
*4.9.4 Чтение изображений* ..... 154
*4.9.5 Изменение размера* ..... 157
4.10 Сохранение/восстановление модели ..... 160
*4.10.1 Сохранение* ..... 160
*4.10.2 Восстановление* ..... 160
Резюме ..... 161

**Часть II БАЗОВЫЙ ШАБЛОН КОНСТРУИРОВАНИЯ** ..... 163

*5 Шаблон процедурного конструирования* ..... 165
5.1 Базовая нейросетевая архитектура ..... 167
5.2 Стержневой компонент ..... 169
*5.2.1 VGG* ..... 169
*5.2.2 ResNet* ..... 171
*5.2.3 ResNeXt* ..... 176
*5.2.4 Xception* ..... 178
5.3 Предстержень ..... 179
5.4 Ученический компонент ..... 180
*5.4.1 ResNet* ..... 182
*5.4.2 DenseNet* ..... 185
5.5 Задачный компонент ..... 187
*5.5.1 ResNet* ..... 188
*5.5.2 Многослойный выход* ..... 189
*5.5.3 SqueezeNet* ..... 192
5.6 За пределами компьютерного зрения: обработка естественного языка ..... 194
*5.6.1 Понимание естественного языка* ..... 194
*5.6.2 Трансформерная архитектура* ..... 196
Резюме ..... 197

*6 Широкие сверточные нейронные сети* ..... 199
6.1 Inception v1 ..... 201
*6.1.1 Нативный модуль Inception* ..... 201
*6.1.2 Модуль Inception v1* ..... 204
*6.1.3 Стержень* ..... 207
*6.1.4 Ученик* ..... 207

*6.1.5 Вспомогательные классификаторы* ..... 208
*6.1.6 Классификатор* ..... 210
6.2 Inception v2: разложение сверток ..... 211
6.3 Inception v3: модернизация архитектуры ..... 214
*6.3.1 Группы и блоки архитектуры Inception* ..... 215
*6.3.2 Нормальная свертка* ..... 219
*6.3.3 Пространственно разделяемая свертка* ..... 220
*6.3.4 Модернизация и имплементация стержня* ..... 220
*6.3.5 Вспомогательный классификатор* ..... 222
6.4 ResNeXt: широкие остаточные нейронные сети ..... 223
*6.4.1 Блок ResNeXt* ..... 224
*6.4.2 Архитектура ResNeXt* ..... 227
6.5 Широкая остаточная сеть ..... 228
*6.5.1 Архитектура WRN-50-2* ..... 228
*6.5.2 Широкий остаточный блок* ..... 229
6.6 За пределами компьютерного зрения: структурированные данные ..... 230
Резюме ..... 233

**7 *Альтернативные шаблоны связности*** ..... 235
7.1 DenseNet: плотносвязанная сверточная нейронная сеть ..... 237
*7.1.1 Плотная группа* ..... 237
*7.1.2 Плотный блок* ..... 240
*7.1.3 Макроархитектура DenseNet* ..... 243
*7.1.4 Плотный переходный блок* ..... 244
7.2 Xception: экстремальное начало ..... 245
*7.2.1 Архитектура Xception* ..... 247
*7.2.2 Входной поток Xception* ..... 249
*7.2.3 Срединный поток модели Xception* ..... 252
*7.2.4 Выходной поток архитектуры Xception* ..... 253
*7.2.5 Свертка, разделяемая по глубине* ..... 256
*7.2.6 Свертка вглубь* ..... 256
*7.2.7 Точечная свертка* ..... 256
7.3 SE-Net: сдавливание и возбуждение ..... 258
*7.3.1 Архитектура SE-Net* ..... 258
*7.3.2 Группа и блок архитектуры SE-Net* ..... 259
*7.3.3 Связь SE* ..... 261
Резюме ..... 262

**8 *Мобильные сверточные нейронные сети*** ..... 263
8.1 MobileNet v1 ..... 264
*8.1.1 Архитектура* ..... 265
*8.1.2 Множитель ширины* ..... 266
*8.1.3 Множитель разрешающей способности* ..... 267
*8.1.4 Стержень* ..... 268
*8.1.5 Ученик* ..... 271
*8.1.6 Классификатор* ..... 273
8.2 MobileNet v2 ..... 274
*8.2.1 Архитектура* ..... 275

*8.2.2 Стержень* ....276
*8.2.3 Ученик* ....277
*8.2.4 Классификатор* ....281
8.3 SqueezeNet ....282
*8.3.1 Архитектура* ....283
*8.3.2 Стержень* ....284
*8.3.3 Ученик* ....285
*8.3.4 Классификатор* ....288
*8.3.5 Обходные соединения* ....290
8.4 ShuffleNet v1 ....294
*8.4.1 Архитектура* ....295
*8.4.2 Стержень* ....295
*8.4.3 Ученик* ....296
8.5 Развертывание ....304
*8.5.1 Квантизация* ....304
*8.5.2 Конверсия и предсказание с TF Lite* ....306
Резюме ....308

9 ***Автокодировщики*** ....309
9.1 Глубокие нейросетевые автокодировщики ....310
*9.1.1 Архитектура автокодировщика* ....310
*9.1.2 Кодировщик* ....312
*9.1.3 Декодировщик* ....313
*9.1.4 Тренировка* ....313
9.2 Сверточные автокодировщики ....315
*9.2.1 Архитектура* ....316
*9.2.2 Кодировщик* ....317
*9.2.3 Декодировщик* ....318
9.3 Разреженные автокодировщики ....320
9.4 Автокодировщики для устранения шума ....321
9.5 Сверхразрешающая способность ....322
*9.5.1 Сверхразрешение на основе предотбора с повышенной частотой* ....323
*9.5.2 Сверхразрешение на основе постотбора с повышенной частотой* ....326
9.6 Предложные задачи ....330
9.7 За пределами компьютерного зрения: последовательность к последовательности ....333
Резюме ....334

**Часть III РАБОТА С КОНВЕЙЕРАМИ** ....336

10 ***Гиперпараметрическая настройка*** ....338
10.1 Инициализация весов ....340
*10.1.1 Распределения весов* ....341
*10.1.2 Лотерейная гипотеза* ....342
*10.1.3 Разминка (численная стабилизация)* ....344
10.2 Основы гиперпараметрического поиска ....347

*10.2.1 Ручной метод гиперпараметрического поиска* ..... 349
*10.2.2 Решеточный поиск* ..... 350
*10.2.3 Случайный поиск* ..... 351
*10.2.4 Инструмент настройки KerasTuner* ..... 354
10.3 Планировщик скорости усвоения ..... 357
*10.3.1 Параметр затухания в Keras* ..... 357
*10.3.2 Планировщик скорости усвоения в Keras* ..... 358
*10.3.3 Рампа* ..... 359
*10.3.4 Постоянный шаг* ..... 360
*10.3.5 Косинусное закаливание* ..... 361
10.4 Регуляризация ..... 364
*10.4.1 Регуляризация весов* ..... 364
*10.4.2 Сглаживание меток* ..... 365
10.5 За пределами компьютерного зрения ..... 367
Резюме ..... 368

11 ***Перенос обучения*** ..... 369
11.1 Предварительно построенные модели TF.Keras ..... 371
*11.1.1 Базовая модель* ..... 372
*11.1.2 Преднатренированные на ImageNet модели для предсказаний* ..... 374
*11.1.3 Новый классификатор* ..... 375
11.2 Предварительно построенные модели TF Hub ..... 380
*11.2.1 Применение преднатренированных моделей TF Hub* ..... 381
*11.2.2 Новый классификатор* ..... 383
11.3 Перенос обучения между предметными областями ..... 385
*11.3.1 Похожие задачи* ..... 385
*11.3.2 Несовпадающие задачи* ..... 387
*11.3.3 Предметно-специфичные веса* ..... 390
*11.3.4 Инициализация предметно-переносимыми весами* ..... 392
*11.3.5 Отрицательный перенос* ..... 394
11.4 За пределами компьютерного зрения ..... 394
Резюме ..... 395

12 ***Распределения данных*** ..... 396
12.1 Типы распределений ..... 397
*12.1.1 Популяционное распределение* ..... 398
*12.1.2 Выборочное распределение* ..... 399
*12.1.3 Подпопуляционное распределение* ..... 401
12.2 Вне распространения ..... 402
*12.2.1 Курируемый набор данных MNIST* ..... 402
*12.2.2 Настройка среды* ..... 403
*12.2.3 Серьезное испытание («дикой природой»)* ..... 404
*12.2.4 Тренировка в качестве глубокой нейросети* ..... 405
*12.2.5 Тренировка в качестве сверточной нейросети* ..... 412
*12.2.6 Обогащение изображений* ..... 415
*12.2.7 Заключительный тест* ..... 418
Резюме ..... 419

13 *Конвейер данных* ..... 420
13.1 Форматы и хранение данных ..... 422
*13.1.1 Форматы сжатых и сырых изображений* ..... 423
*13.1.2 Формат HDF5* ..... 427
*13.1.3 Формат DICOM* ..... 432
*13.1.4 Формат TFRecord* ..... 434
13.2 Подача данных ..... 440
*13.2.1 NumPy* ..... 441
*13.2.2 TFRecord* ..... 443
13.3 Предобработка данных ..... 446
*13.3.1 Предобработка с помощью предстержня* ..... 446
*13.3.2 Предобработка с помощью расширенного TensorFlow (TF Extended)* ..... 455
13.4 Обогащение данных ..... 460
*13.4.1 Инвариантность* ..... 461
*13.4.2 Обогащение с помощью tf.data* ..... 464
*13.4.3 Предстержень* ..... 465
Резюме ..... 466

14 *Конвейер тренировки и развертывания* ..... 467
14.1 Подача данных в модель ..... 469
*14.1.1 Подача данных в модель с помощью tf.data.Dataset* ..... 474
*14.1.2 Распределенная подача с помощью tf.Strategy* ..... 478
*14.1.3 Подача данных в модель с помощью TFX* ..... 480
14.2 Планировщики тренировки ..... 488
*14.2.1 Версионирование конвейера* ..... 490
*14.2.2 Метаданные* ..... 492
*14.2.3 История* ..... 494
14.3 Оценивание моделей ..... 496
*14.3.1 Кандидатная модель против одобренной модели* ..... 496
*14.3.2 Оценивание в TFX* ..... 501
14.4 Обслуживание предсказательных запросов ..... 504
*14.4.1 Обслуживание по требованию (в реальном времени)* ..... 505
*14.4.2 Пакетное предсказание* ..... 508
*14.4.3 Конвейерные компоненты TFX для развертывания* ..... 510
*14.4.4 A/B-тестирование* ..... 512
*14.4.5 Балансировка нагрузки* ..... 514
*14.4.6 Непрерывное оценивание* ..... 516
14.5 Эволюция в конструировании производственных конвейеров ..... 517
*14.5.1 Машинное обучение в качестве конвейера* ..... 518
*14.5.2 Машинное обучение как производственный процесс CI/CD* ..... 519
*14.5.3 Консолидация моделей в производстве* ..... 519
Резюме ..... 521

*Предметный указатель* ..... 522

# *Предисловие*

Одна из моих обязанностей в качестве сотрудника Google состоит в том, чтобы обучать инженеров-программистов приемам применения машинного обучения. У меня уже был опыт создания онлайновых учебных занятий, встреч, презентаций на конференциях, рабочих семинаров и курсовых работ для частных школ программирования и аспирантур университетов, но я всегда ищу новые способы эффективного преподавания.

До Google я в течение 20 лет проработал в японской информационно-технологической индустрии в качестве главного научного сотрудника – и все время без глубокого обучения. Почти все, что я вижу сегодня, мы делали в инновационных лабораториях 15 лет назад; разница лишь в том, что нам нужен был коллектив, полный ученых, и огромный бюджет. Невероятно, как все так быстро поменялось в результате повсеместного внедрения технологии глубокого обучения.

Еще в конце 2000-х годов я работал с небольшими структурированными наборами геопространственных данных из национальных и международных источников, разбросанных по всему миру. Коллеги называли меня исследователем данных, но никто не знал, что это такое на самом деле. Затем появились большие данные, которые проявили мою неосведомленность об инструментах и каркасах больших данных, и я неожиданно перестал быть исследователем данных. Вот незадача. Мне пришлось поднапрячься и изучить инструменты и концепции, лежащие в основе больших данных, и я снова стал исследователем данных.

И вот появилось машинное обучение на больших наборах данных, такое как линейная/логистическая регрессия и анализ CART, а я не использовал статистику со времен аспирантуры десятилетней давности, и я снова перестал быть исследователем данных. Вот дела! Мне пришлось поднапрячься и выучить статистику заново, и я снова стал исследователем данных. Затем пришло глубокое обучение, а я не знал теории и основ нейронных сетей, и я внезапно перестал быть исследователем данных. Что опять? Но я снова поднапрягся и изучил теорию и другие основы глубокого обучения. И опять-таки, теперь я – снова исследователь данных.

# *Признательности*

Хотел бы поблагодарить всех сотрудников издательства Manning, которые помогали на протяжении всего этого процесса. Франческу Лефковиц, редактора по разработке; Дейдруе Хиам, редактора проектов; Шарон Уилки, редактора-копирайтера; Кери Хейлз, корректора; и Александра Драгошавлевича, редактора-рецензента.

Всем рецензентам: Ариэль Гамино, Арне Питер Раульф, Барри Сигел, Брайан Р. Гейнс, Кристофер Маршалл, Кертис Бейтс, Эрос Педрини, Хильде Ван Гизель, Ишан Хурана, Джен Ли, Картикеяраджан Раджендран, Майкл Кареев, Мухаммад Сохаиб Ариф, Ник Васкес, Нинослав Черкез, Оливер Кортен, Пиюш Мехта, Ричард Тобиас, Ромит Синг-хай, Саяк Пол, Серджио Говони, Симона Сгуацца, Удендран Мудалияр, Вишвеш Рави Шримали и Витон Витанис, – ваши предложения помогли сделать эту книгу лучше.

Всем сотрудникам Google Cloud AI, которые поделились своими личными знаниями и сведениями в области клиентских предпочтений, – ваши идеи помогли книге охватить более широкую аудиторию.

# Об этой книге

## Кому следует прочитать эту книгу

Добро пожаловать в мое последнее начинание – книгу «Шаблоны и практика глубокого обучения». Эта книга предназначена для инженеров-программистов, инженеров машинного обучения, а также младших, средних и старших специалистов-исследователей данных. Хотя вы, возможно, посчитаете, что начальные главы будут полезны для последней группы, мой уникальный подход, скорее всего, даст вам дополнительную информацию и поможет освежить знания. Книга построена так, чтобы каждый читатель достиг точки «зажигания» и смог продвигаться вперед к глубокому обучению самостоятельно.

Я преподаю шаблоны конструирования и образцы практики главным образом в контексте компьютерного зрения, так как именно здесь шаблоны конструирования появились для глубокого обучения впервые. Разработки в области понимания естественного языка и моделей структурированных данных отставали и по-прежнему были сосредоточены на классических подходах. Но по мере того, как они догоняли, эти области вырабатывали свои собственные шаблоны конструирования для решения задач глубокого обучения, и я излагаю эти шаблоны и образцы практики на протяжении всей книги.

Несмотря на то что я демонстрирую исходный код моделей компьютерного зрения, мое внимание сосредоточено на концепциях, лежажих в основе подходов и инноваций: тому, как они устроены и почему они устроены таким образом. Указанные опорные концепции применимы к обработке естественного языка, структурированным данным, обработке сигналов и другим областям, и, резюмируя, вы должны быть в состоянии адаптировать эти концепции, методы и образцы практики к задачам в вашей предметной области. Многие модели и методы, которые я обсуждаю, не зависят от предметной области, и на протяжении всей книги, где это уместно, я также обсуждаю ключевые инновации в областях обработки естественного языка, понимания естественного языка и структурированных данных.

Если говорить об общей подготовке, то вы должны знать, по крайней мере, основы Python. Все в порядке, если вы все еще пытаетесь разобраться в том, что такое включение в список или генератор, или если у вас все еще есть некоторая путаница в отношении странного среза многомерного массива и того, какие объекты являются мутируемыми и немутируемыми в куче. Для этой книги данного уровня будет достаточно.

Какой должна быть подготовка тех инженеров-программистов, которые хотят стать инженерами машинного обучения? Инженер машинного обучения (MLE) – это инженер-прикладник. Вам не требуется знать статистику (реально не нужно!), и вам не требуется знать теорию вычислений. Если вы заснули в колледже на уроке математики на теме производной, то это нормально, и если кто-то попросит вас выполнить матричное умножение, не стесняйтесь спрашивать, зачем оно нужно.

Ваша задача состоит в том, чтобы изучить «кнопки и рычаги» вычислительного каркаса и применять свои навыки и опыт для выработки решений реально существующих задач. Вот в чем я собираюсь вам помочь, и вот в чем суть шаблонов конструирования с использованием модуля TF.Keras.

Эта книга предназначена для инженеров машинного обучения и исследователей данных на сопоставимых уровнях. Тем же, кто следует по пути анализа данных, я рекомендую изучить дополнительные материалы, связанные со статистикой.

Прежде чем мы начнем, я хочу объяснить, как вы будете учиться, поэтому в данном первом разделе больше рассказывается о моей философии и подходе к преподаванию. Затем мы рассмотрим некоторые основополагающие материалы, включая терминологию, переход от классического или семантического ИИ к узкому или статистическому ИИ, а также проведем обзор основных шагов машинного обучения. Наконец, мы подробно остановимся на том, чему посвящена книга: на современном подходе к машинному обучению, основанному на *консолидации моделей*.

Я не использую традиционный западный подход: заучивание наизусть, повторение, повторение, проверка правильности ответов и затем продвижение вертикально вверх. Помимо моего мнения о том, что такой подход к преподаванию менее эффективен, я считаю, что он непреднамеренно дискриминирует учащихся.

Вместо этого я имел возможность преподавать инженерное дело и естественные науки в различных культурах и методиках преподавания и разработал уникальный стиль преподавания, с привлечением того, что я называю *боковым подходом*: я начинаю с ключевых понятий, а затем продвигаюсь по спирали, используя то, что я называю *абстракцией*. Когда начнут задаваться вопросы, я постепенно перехожу к указыванию другим студентам на их мысли по поводу ответов на эти вопросы, а потом размышляю над их мыслями. Я не провожу контрольные работы, в которых студенты пытаются получить 100 %. Вместо этого я даю задания, которые каждый студент провалит. Я по-

зволяю студентам биться над задачей изо всех сил, и при этом они начинают открывать для себя опорные принципы того, что им нужно усвоить. Например, я могу дать задание натренировать стандартную модель ResNet50 с использованием набора данных CIFAR-10, отметив, что авторы соответствующих статей по ResNet достигли на CIFAR-10 точности 97 %. Каждый студент провалит решение, модель не сойдется, они не наберут более 70 % и так далее.

Затем я собираю студентов в группы, чтобы решать задачи вместе. Совещаясь друг с другом, они учатся делать совместные обобщения. И прежде чем они достаточно созреют, я совершаю прыжок, в котором ставлю перед студентами еще одну трудноразрешимую задачу, – и процесс начинается снова. Я никогда не даю студентам возможности заучивать наизусть.

Используя свой пример, я могу разместить на доске четыре возможных решения, например: 1) обогащение изображений, 2) еще больше регуляризации, 3) еще больше гиперпараметрического поиска, 4) отложить снижение размера изображения глубоко внутрь нейронной сети (это правильный ответ). Далее в середине я останавливаю студентов и прошу каждую группу указать опробованное ими решение и то, что они усвоили к тому моменту. Затем я объясняю причины правильности/неправильности каждого решения, а потом снова меняю задачу.

По мере того как студенты переходят на более продвинутые уровни, я переключаюсь с роли учителя на то, что я называю ролью магистранта, и участвую в обучении. Студенты учат меня и друг друга так же, как я учу их. Я наблюдаю за каждым студентом и ищу то, что я называю *зажиганием*, – этап, когда студент начнет саморазвиваться как ученик, то есть когда он учится постоянно. В своем методе преподавания я замечаю, что весь класс собирается вместе и ни один ученик не остается позади.

Время от времени на одно из моих занятий приходил администратор школы программирования. Он слышал доносящуюся от студентов болтовню и хотел понаблюдать за тем, как это работает. Разумеется, администраторы испытывают потребность в том, чтобы всему давать название. В одной частной школе программирования администратор описал мою методику как «каждый становится учеником». Студенты учатся у учителя, учитель учится у студентов, и студенты учатся у студентов. Администратор назвал ее «Давайте учиться вместе».

У меня же для моей методики преподавания есть собственное название, а именно «Я верю в себя». Я часто говорю своим ученикам: как можно верить в меня («учителя»), если вы прежде всего не верите в себя?

## Как эта книга организована: дорожная карта

Эта книга состоит из трех частей: основы, общие шаблоны конструирования и шаблоны конструирования для решения задач тренировки и развертывания в производстве.

Часть I «Основы глубокого обучения» предоставляет читателям обновленную информацию о глубоком обучении, которая включает введение в сверточные нейронные сети, а также обсуждение концепций и терминологии, которые являются сегодня магистральными для всех областей – компьютерного зрения, обработки естественного языка и структурированных данных.

Шаблоны конструирования моделей представлены в части II «Базовые шаблоны конструирования». В главах 5–7 я ввожу современные шаблоны конструирования и способы их применения ко многим современным и некогда передовым моделям глубокого обучения. Я расскажу о шаблоне процедурного реиспользования, который был преобладающим подходом для моделей, которые конструировались вручную. Я излагаю подходы к конструированию, усовершенствования и плюсы/минусы крупных моделей, обнаруженных исследователями для движения послойно вглубь (глава 5), движения послойно вширь (глава 6), и применению альтернативных или готовых («прямо из коробки») шаблонов связности (глава 7).

В главе 5 рассматривается шаблон процедурного конструирования для сверточных нейронных сетей, а также разработка остаточных блоков с отождествляющими связями с вниманием в трансформерах для понимания естественного языка.

В главе 6 подробно рассказывается о шаблоне процедурного конструирования для сверточных нейронных сетей и о том, как исследователи разведали движение послойно в ширину в качестве альтернативы движению в глубину. Я показываю, каким образом такой подход, как ResNeXt, привел к достижению сравнимой точности в сопоставлении с глубокими слоями с меньшей подверженностью запоминанию и исчезающим градиентам. Я также проведу разведывательный анализ степени релевантности широких сверточных нейронных сетей для разработок в широких и глубоких моделях TabNet для структурированных данных.

В главе 7 рассматриваются шаблоны конструирования моделей. Указанные шаблоны разведывают другие альтернативные соединения между слоями, чтобы двигаться послойно вглубь либо вширь с целью повышения точности, сокращения числа параметров и увеличения прироста информации в промежуточном латентном пространстве внутри модели.

В главе 8 инспектируются уникальные конструктивные соображения и особые ограничения для мобильных сверточных нейронных сетей. Из-за ограничений этих устройств по памяти необходимо учитывать компромиссы между размером и точностью. Я расскажу о прогрессе в этих компромиссах, плюсах/минусах и о том, как конструкции мобильных сетей отличаются от их крупномодельных визави, чтобы учесть эти компромиссы.

В главе 9 представлены автокодировщики для неконтролируемого обучения. Практическая применимость автокодировщиков в качестве автономных моделей  очень узка. Но сделанные автокодировщи-

ками открытия способствовали прогрессу в предтренировке моделей. Такие модели лучше обобщают на обслуживание запросов вне распределения, то есть на предсказательные запросы к модели, развернутой в производственной среде, которые имеют иное распределение, чем данные, на которых модель была натренирована. Я также разведаю сопоставимость автокодировщиков с векторными вложениями в отрасли понимания естественного языка.

Все модели во второй части этой книги внесли эпохальный вклад в исследования и разработки в области глубокого обучения и продолжают использоваться сегодня, либо их вклад был включен в современные модели.

В части III «Работа с конвейерами» рассматриваются шаблоны конструирования и образцы практики для производственных конвейеров. В главе 10 мы рассмотрим гиперпараметрическую настройку, как ручную, так и автоматическую. Я изложу конструктивные решения, плюсы/минусы и лучшие практические приемы определения пространства поиска и шаблонов поиска в нем.

В главе 11 обсуждается тема переноса обучения (трансферного обучения) и вводятся концепции и методы манипулирования переносом весов и настройки под аналогичные и отдаленные задачи. Я также рассматриваю приложение по переносу предметных знаний с целью реиспользования весов во время предтренировки; данное приложение предназначено для моделей, которые тренируются полностью с нуля.

В главах с 12 по 14 выполняется высокоуровневый обзор производственных конвейеров. В главах 12 и 13 мы погрузимся в эту тему со стороны данных. Глава 12, в которой рассказывается о распределении данных, является единственной, в которой статистика излагается подробно. С 2017 года, когда можно было ожидать, что специалист будет обладать знаниями в области статистики на уровне доктора философии, многое изменилось. Сегодня многое из того скрыто или же автоматизировано в каркасах глубокого обучения, таких как TensorFlow. Понимание распределения данных и пространства поиска остается одной из преобладающих областей ожидаемых знаний из области статистики, и оно может существенно влиять на стоимость тренировки и способность модели обобщать после ее развертывания в производстве.

Наконец, в главах 13 и 14 мы переходим со стороны данных на сторону развертывания. Я рассказываю о концепциях и лучших образцах практики конструирования со стороны данных, а затем со стороны тренировки производственного конвейера.

## Об исходном коде

Эта книга содержит массу примеров исходного кода как в пронумерованных листингах, так и во вставках, встроенных в обычный текст. В обоих случаях исходный код отформатирован шрифтом фиксированной

ширины, чтобы отделять его от обычного текста. Иногда исходный код также выделяется **жирным шрифтом**, чтобы выделять исходный код, который изменился по сравнению с предыдущими шагами в главе, например когда в существующую строку исходного кода добавляется новая функция.

Во многих случаях исходный код был переформатирован; мы добавили разрывы строк и переработали отступы, чтобы уложиться в доступное пространство книжной страницы. Кроме того, комментарии в исходном коде часто из листингов удалялись, когда исходный код описывался в тексте. Многочисленные листинги сопровождаются аннотациями исходного кода, выделяя важные концепции.

Все приводимые в книге примеры исходного кода написаны на Python и являются рабочими; правда, в них могут отсутствовать инструкции импорта. Во многих случаях образцы исходного кода являются частью более крупного компонента, такого как модель. В подобных случаях весь исходный код доступен в моем публичном репозитории для связей с разработчиками Google Cloud AI на GitHub (https://github.com/GoogleCloudPlatform/keras-idiomatic-programmer/tree/master/zoo).

## Другие онлайновые ресурсы

Я использую вычислительный каркас TensorFlow 2.x, в который включен API моделей Keras. Я думаю, что сочетание этих двух факторов является фантастическим средством для образования, выходящим за рамки их производственной ценности.

Материал книги является мультимодальным. В дополнение к книге и полным образцам исходного кода в репозитории на моем YouTube-канале для связей с разработчиками Google Cloud AI (www.youtube.com/канал/uc8ov0vkzhtp8_puwedzlbjg) есть слайды презентаций, семинары, лабораторные работы и предварительно записанные лекции по каждой главе.

## Отзывы и пожелания

Мы всегда рады отзывам наших читателей. Расскажите нам, что вы думаете об этой книге, – что понравилось или, может быть, не понравилось. Отзывы важны для нас, чтобы выпускать книги, которые будут для вас максимально полезны.

Вы можете написать отзыв на нашем сайте www.dmkpress.com, зайдя на страницу книги и оставив комментарий в разделе «Отзывы и рецензии». Также можно послать письмо главному редактору по адресу dmkpress@gmail.com; при этом укажите название книги в теме письма.

Если вы являетесь экспертом в какой-либо области и заинтересованы в написании новой книги, заполните форму на нашем сайте по

адресу http://dmkpress.com/authors/publish_book/ или напишите в издательство по адресу dmkpress@gmail.com.

## Список опечаток

Хотя мы приняли все возможные меры для того, чтобы обеспечить высокое качество наших текстов, ошибки все равно случаются. Если вы найдете ошибку в одной из наших книг, мы будем очень благодарны, если вы сообщите о ней главному редактору по адресу dmkpress@gmail.com. Сделав это, вы избавите других читателей от недопонимания и поможете нам улучшить последующие издания этой книги.

## Нарушение авторских прав

Пиратство в интернете по-прежнему остается насущной проблемой. Издательства «ДМК Пресс» и Manning Publications очень серьезно относятся к вопросам защиты авторских прав и лицензирования. Если вы столкнетесь в интернете с незаконной публикацией какой-либо из наших книг, пожалуйста, пришлите нам ссылку на интернет-ресурс, чтобы мы могли применить санкции.

Ссылку на подозрительные материалы можно прислать по адресу электронной почты dmkpress@gmail.com.

Мы высоко ценим любую помощь по защите наших авторов, благодаря которой мы можем предоставлять вам качественные материалы.

# *Об авторе*

Я твердо верю, что мой жизненный опыт делает меня одним из самых идеальных людей для преподавания концепций глубокого обучения. Когда эта книга выйдет из печати в первый раз, мне будет почти 60 лет. Я обладаю богатыми знаниями и опытом, которые сегодня соответствуют ожиданиям сотрудников. В 1987 году я получил ученую степень в области искусственного интеллекта. Я специализировался на обработке естественного языка. Когда я закончил колледж, то думал, что буду писать говорящие книги. Как оказалось, это было время зимы искусственного интеллекта.

В начале своей карьеры я выбрал другие направления. Прежде всего я стал экспертом в области государственной безопасности для мейнфреймов. По мере того как я набирался все больше опыта в конструировании и программировании ядер операционных систем, я стал разработчиком ядра для UNIX, будучи одним из авторов современного тяжеловесного ядра UNIX. В те же годы я участвовал в популяризации свободно распространяемого ПО «shareware» (еще до раскрытия исходного кода) и был основателем WINNIX, условно-бесплатной программы, которая конкурировала с коммерческим инструментарием MKS для исполнения оболочки UNIX и команд в среде DOS.

Впоследствии я разработал низкоуровневый инструментарий объектного кода. В начале 1990-х я стал экспертом как в области вычислений на защищенном уровне, так и в области компиляторов/ассемблеров для массово-параллельных вычислений. Я разработал инструмент MetaC, который обеспечивал инструментальную поддержку ядер операционных систем как традиционных операционных систем, так и высокозащищенных и массово-параллельных компьютеров.

В конце 1990-х годов я сменил карьеру и стал научным сотрудником японской корпорации Sharp. Через пару лет я стал главным научным сотрудником этой компании в Северной Америке. За 20-летний период Sharp подала более 200 заявок на патенты в США на мои исследования, из которых 115 были удовлетворены. Мои патенты охватывали области солнечной энергетики, телеконференций, ви-

зуализации, цифровых интерактивных вывесок и автономных транспортных средств. Кроме того, в 2014–2015 годах я был признан ведущим мировым экспертом по открытым данным и онтологиям данных и основал организацию opengeocode.

В марте 2017 года, по настоянию моего друга, я решил посмотреть, «что это за диковинка такая, которую называют глубоким обучением». Для меня это было естественно. У меня был большой опыт работы с данными, я работал специалистом и исследователем по обработке изображений, имел степень магистра искусственного интеллекта, работал над автономными транспортными средствами – все это, казалось, укладывалось в одну линию. И стало быть, я совершил прыжок.

Летом 2018 года Google обратилась ко мне с просьбой стать сотрудником Google Cloud AI. Я принял должность в октябре того же года. Это был и остается великолепный опыт работы в Google. Сегодня я работаю с огромным числом экспертов в области искусственного интеллекта как в Google, так и с корпоративными клиентами Google, обучая, наставляя, консультируя и решая задачи по внедрению глубокого обучения в больших производственных масштабах.

# *Об иллюстрации на обложке*

Рисунок на обложке книги «Шаблоны и практика глубокого обучения» озаглавлен «Индиец», или человек родом из Индии. Иллюстрация взята из коллекции костюмов из разных стран Жака Грассе де Сен-Совера (1757–1810) под названием «Костюмы разных стран», опубликованной во Франции в 1784 году. Каждая иллюстрация тщательно прорисована и оформлена вручную. Богатое разнообразие коллекции Грассе де Сен-Совера живо напоминает нам о том, насколько культурно обособленными были города и регионы мира всего 200 лет назад. Изолированные друг от друга люди говорили на разных диалектах и языках. На улицах городов или в деревнях было легко просто по их одежде определить, где они живут и каково их ремесло или положение в жизни.

С тех пор наша манера одеваться изменилась, и региональное разнообразие, столь богатое в то время, исчезло. В настоящее время трудно отличить жителей разных континентов, не говоря уже о разных городах, регионах или странах. Возможно, мы променяли культурное разнообразие на более разнообразную личную жизнь – безусловно, на более разнообразную и быстро развивающуюся технологическую жизнь.

Сегодня, когда трудно отличить одну компьютерную книгу от другой, издательство Manning демонстрирует изобретательность и инициативу компьютерного бизнеса, предлагая обложки книг, основанные на богатом разнообразии региональной жизни двухвековой давности, оживленной картинами Грассе де Сен-Совера.