

УДК 004.85
ББК 32.971.3
Ф43

Ферлитш Э.

Ф43 Шаблоны и практика глубокого обучения / пер. с англ. А. В. Логунова. – М.: ДМК Пресс, 2022. – 538 с.: ил.

ISBN 978-5-93700-113-9

В книге рассматриваются актуальные примеры создания приложений глубокого обучения с учетом десятилетнего опыта работы автора в этой области. Вы сэкономите часы проб и ошибок, воспользовавшись представленными здесь шаблонами и приемами. Проверенные методики, образцы исходного кода и блестящий стиль повествования позволят с увлечением освоить даже непростые навыки. По мере чтения вы получите советы по развертыванию, тестированию и техническому сопровождению ваших проектов.

Издание предназначено для инженеров машинного обучения, знакомых с Python и глубоким обучением.

УДК 004.85
ББК 32.971.3

Original English language edition published by Manning Publications USA. Russian-language edition copyright © 2022 by DMK Press. All rights reserved.

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.

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

<i>Предисловие</i>	13
<i>Признательности</i>	14
<i>Об этой книге</i>	15
<i>Об авторе</i>	22
<i>Об иллюстрации на обложке</i>	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

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

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 предтбора с повышенной частотой	323
9.5.2	Сверхразрешение на основе посттбора с повышенной частотой	326
9.6	Предлоговые задачи	330
9.7	За пределами компьютерного зрения: последовательность к последовательности	333
	Резюме	334

Часть III РАБОТА С КОНВЕЙЕРАМИ

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