

УДК 004.032.2
ББК 32.972.1
Г19

Ганегедара Т.

Г19 Обработка естественного языка с TensorFlow / пер. с англ. В. С. Яценкова. – М.: ДМК Пресс, 2020. – 382 с.: ил.

ISBN 978-5-97060-756-5

TensorFlow – библиотека на языке Python для реализации систем глубокого обучения, позволяющих решать в том числе уникальные задачи по обработке естественного языка.

Автор книги излагает общие принципы работы NLP и построения нейронных сетей, описывает стратегии обработки больших объемов данных, а затем переходит к практическим темам. Вы узнаете, как использовать технологию Word2vec и ее расширения для создания представлений, превращающих последовательности слов в числовые векторы, рассмотрите примеры решения задач по классификации предложений и генерации текста, научитесь применять продвинутые рекуррентные модели и сможете самостоятельно создать систему нейронного машинного перевода.

Издание предназначено для разработчиков, которые, используя лингвистические данные, применяют и совершенствуют методы машинной обработки естественного языка.

УДК 004.032.2
ББК 32.972.1

Original English language edition published by Packt Publishing Ltd., UK. Copyright © 2018 Packt Publishing. Russian-language edition copyright © 2020 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.

ISBN 978-1-78847-831-1 (англ.)
ISBN 978-5-97060-756-5 (рус.)

Copyright © 2018 Packt Publishing
© Оформление, издание, перевод, ДМК Пресс, 2020

Содержание

Об авторе	13
О рецензентах	14
Предисловие	15
Глава 1. Введение в обработку естественного языка	21
Что такое обработка естественного языка?	21
Задачи обработки естественного языка	22
Традиционный подход к обработке естественного языка	24
Подробности традиционного подхода	24
Недостатки традиционного подхода	29
Революция глубокого обучения в обработке естественного языка	30
История глубокого обучения	30
Современное состояние глубокого обучения и NLP	32
Устройство простой глубокой модели – полносвязной нейронной сети	33
Что вы найдете дальше в этой книге?	34
Знакомство с рабочими инструментами	38
Обзор основных инструментов	38
Установка Python и scikit-learn	39
Установка Jupyter Notebook	39
Установка TensorFlow	40
Заключение	41
Глава 2. Знакомство с TensorFlow	42
Что такое TensorFlow?	42
Начало работы с TensorFlow	43
Подробно о клиенте TensorFlow	45
Архитектура TensorFlow – что происходит при запуске клиента?	46
Кафе Le TensorFlow – пояснение устройства TensorFlow на примере	49
Входные данные, переменные, выходные данные и операции	49
Определение входных данных в TensorFlow	50
Объявление переменных в TensorFlow	55
Объявление выходных данных TensorFlow	57
Объявление операций TensorFlow	57
Повторное использование переменных с областью видимости	66
Реализация нашей первой нейронной сети	68
Подготовка данных	68
Определение графа TensorFlow	69
Запуск нейронной сети	71
Заклучение	72

Глава 3. Word2vec и вектор слова в пространстве смыслов	74
Что такое представление и значение слова?	75
Классические подходы к представлению слов	76
Внешняя лексическая база знаний WordNet для изучения представлений слов	76
Прямое унитарное кодирование	79
Метод TF-IDF	80
Матрица совместной встречаемости	81
Word2vec – нейросетевой подход к изучению представления слова	82
Упражнение: королева = король – он + она?	83
Разработка функции потери для изучения представлений слов	87
Алгоритм skip-gram	87
От необработанного текста до структурированных данных	88
Изучение представлений слов с помощью нейронной сети	88
Реализация алгоритма skip-gram с TensorFlow	98
Алгоритм CBOW	100
Реализация алгоритма CBOW с TensorFlow	100
Заключение	102
Глава 4. Углубленное изучение Word2vec	103
Исходный алгоритм skip-gram	103
Реализация исходного алгоритма skip-gram	104
Сравнение исходного и улучшенного алгоритмов skip-gram	106
Сравнение skip-gram и CBOW	107
Сравнение продуктивности	108
Кто же победитель, skip-gram или CBOW?	111
Расширения алгоритмов представления слов	113
Использование униграммного распределения для отрицательной выборки	113
Реализация отрицательной выборки на основе униграмм	113
Подвыборка – вероятностное игнорирование общих слов	115
Реализация подвыборки	116
Сравнение CBOW и его расширений	116
Более современные алгоритмы, расширяющие skip-gram и CBOW	117
Ограничение алгоритма skip-gram	117
Структурированный алгоритм skip-gram	118
Функция потерь	119
Модель непрерывного окна	120
GloVe – представление на основе глобальных векторов	121
Знакомство с GloVe	121
Реализация алгоритма GloVe	122
Классификация документов с помощью Word2vec	123
Исходный набор данных	124
Классификация документов при помощи представлений слов	125
Реализация – изучение представлений слов	125
Реализация – от представлений слов к представлениям документов	126

Кластеризация документов и визуализация представлений.....	126
Проверка некоторых выбросов.....	126
Кластеризация/классификация документов с К-средним	129
Заключение	130

Глава 5. Классификация предложений с помощью сверточных нейронных сетей.....

Знакомство со сверточными нейронными сетями	132
Основы CNN	133
Возможности сверточных нейросетей.....	135
Устройство сверточных нейросетей.....	136
Операция свертки.....	136
Операция субдискретизации.....	139
Полностью связанные слои.....	141
Собираем CNN из компонентов	142
Упражнение – классификация изображений из набора MNIST	143
Источник данных.....	143
Реализация CNN	143
Анализ прогнозов, сделанных CNN.....	146
Классификация предложений с помощью сверточной нейросети.....	147
Структура нейросети.....	147
Растянутая субдискретизация	150
Реализация классификации предложений	151
Заключение	154

Глава 6. Рекуррентные нейронные сети

Знакомство с рекуррентными нейронными сетями.....	156
Проблема с нейросетью прямого распространения	156
Моделирование с помощью рекуррентных нейронных сетей.....	157
Устройство рекуррентной нейронной сети в деталях	159
Обратное распространение во времени	160
Как работает обратное распространение	160
Почему нельзя использовать простое обратное распространение	161
Обратное распространение во времени и обучение RNN	162
Усеченное обратное распространение во времени.....	163
Ограничения ВРТТ – исчезающие и взрывающиеся градиенты	163
Применение рекуррентных нейросетей.....	165
Один-к-одному	166
Один-ко-многим	166
Многие-к-одному	167
Многие-ко-многим.....	168
Генерация текста с помощью рекуррентной нейросети.....	168
Определение гиперпараметров.....	169
Распространение входов во времени для усеченного ВРТТ.....	169
Определение набора данных для валидации	170
Определение весов и смещений.....	170

Определение переменных состояния	171
Вычисление скрытых состояний и выходов с развернутыми входами	171
Расчет потерь	172
Сброс состояния в начале нового сегмента текста	172
Расчет результата проверки.....	172
Расчет градиентов и оптимизация.....	173
Вывод сгенерированного фрагмента текста.....	173
Оценка качества текста.....	174
Перplexия – измерение качества созданного текста.....	175
Рекуррентные нейронные сети с контекстными признаками.....	176
Особенности устройства RNN-CF	177
Реализация RNN-CF.....	178
Текст, созданный с помощью RNN-CF	183
Заключение	186
Глава 7. Сети с долгой краткосрочной памятью	188
Устройство и принцип работы LSTM	189
Что такое LSTM?.....	189
LSTM в деталях.....	190
Чем LSTM отличаются от стандартных RNN.....	199
Как LSTM решает проблему исчезающего градиента	200
Улучшение LSTM.....	202
Жадная выборка	202
Лучевой поиск.....	203
Использование векторных представлений слов	204
Двунаправленные LSTM (BiLSTM).....	205
Другие варианты LSTM	207
Замочная скважина	207
Управляемые рекуррентные ячейки (GRU)	208
Заключение	210
Глава 8. Применение LSTM для генерации текста	211
Наши данные	211
О наборе данных.....	212
Предварительная обработка данных	214
Реализация LSTM.....	214
Объявление гиперпараметров.....	214
Объявление параметров	215
Объявление ячейки LSTM и ее операций	217
Входные данные и метки	217
Последовательные вычисления для обработки последовательных данных.....	218
Выбор оптимизатора.....	219
Снижение скорости обучения.....	219
Получение прогнозов.....	220
Вычисление перplexии	220

Сброс состояний	221
Жадная выборка против унимодальности	221
Генерация нового текста	221
Пример сгенерированного текста	222
Сравнение качества текстов на выходе разных модификаций LSTM	223
Обычная LSTM-сеть	223
Пример генерации текста при помощи GRU	225
LSTM с замочными скважинами	228
Обучение нейросети и проверка перплексии	230
Модификация LSTM – лучевой поиск	232
Реализация лучевого поиска	232
Пример текста, созданного лучевым поиском	234
Генерация текста на уровне слов вместо n -грамм	235
Проклятие размерности	235
Word2vec спешит на помощь	236
Генерация текста с помощью Word2vec	236
Текст, созданный с помощью LSTM–Word2vec и лучевого поиска	237
Анализ уровня перплексии	239
Использование TensorFlow RNN API	240
Заключение	243

Глава 9. Применение LSTM – генерация подписей к рисункам..... 245

Знакомство с данными	246
Набор данных ILSVRC ImageNet	246
Набор данных MS-COCO	246
Устройство модели для генерации подписей к изображениям	249
Извлечение признаков изображения	250
Реализация – загрузка весов и вывод с помощью VGG-16	252
Создание и обновление переменных	252
Предварительная обработка входов	253
Распространение данных через VGG-16	254
Извлечение векторизованных представлений изображений	255
Прогнозирование вероятностей классов с помощью VGG-16	255
Изучение представлений слов	256
Подготовка подписей для подачи в LSTM	258
Формирование данных для LSTM	259
Определение параметров и процедуры обучения LSTM	260
Количественная оценка результатов	262
BLEU	263
ROUGE	264
METEOR	264
CIDEr	266
Изменение оценки BLEU-4 для нашей модели	267
Подписи, созданные для тестовых изображений	267
Использование TensorFlow RNN API с предварительно обученными векторами слов GloVe	271
Загрузка векторов слов GloVe	271

Очистка данных	272
Использование предварительно изученных представлений с RNN API	274
Заключение	279

Глава 10. Преобразование последовательностей

и машинный перевод	281
Машинный перевод	281
Краткая историческая экскурсия по машинному переводу	282
Перевод на основе правил	282
Статистический машинный перевод (SMT)	284
Нейронный машинный перевод	286
Общие принципы нейронного машинного перевода	288
Устройство NMT	288
Архитектура NMT	289
Подготовка данных для системы NMT	292
Этап обучения	292
Переворачивание исходного предложения	293
Этап тестирования	294
Обучение NMT	294
Вывод перевода из NMT	295
Метрика BLEU – оценка систем машинного перевода	295
Модифицированная точность	296
Штраф за краткость	297
Окончательная оценка BLEU	297
Собственная система NMT с нуля – переводчик с немецкого на английский	297
Знакомство с данными	298
Предварительная обработка данных	298
Изучение представлений слов	299
Кодер и декодер	300
Сквозные вычисления	302
Примеры результатов перевода	304
Обучение NMT одновременно с изучением представлений слов	306
Максимизация совпадений между словарем набора данных и предварительно подготовленными представлениями	306
Объявление слоя представлений как переменной TensorFlow	308
Совершенствование NMT	310
Помощь наставника	310
Глубокие LSTM	312
Механизм внимания	312
Узкое место: вектор контекста	313
Механизм внимания в деталях	314
Результаты работы NMT со вниманием	319
Визуализация внимания к исходным и целевым предложениям	321
Применение моделей Seq2Seq в чат-ботах	322
Обучение чат-бота	322
Оценка чат-ботов – тест Тьюринга	324
Заклучение	324

Глава 11. Современные тенденции и будущее обработки

естественного языка	326
Современные тенденции в NLP	327
Представления слов	327
Нейронный машинный перевод	332
Применение NLP в смежных прикладных областях	334
Сочетание NLP с компьютерным зрением	334
Обучение с подкреплением	336
Генеративные состязательные сети и NLP	338
На пути к искусственному общему интеллекту	340
Обучил одну модель – обучил их все	340
Совместная многозадачная модель – развитие нейронной сети для множества задач NLP	342
NLP для социальных сетей	344
Обнаружение слухов в соцсетях	344
Обнаружение эмоций в социальных сетях	345
Анализ политического наполнения в твитах	345
Новые задачи и вызовы	347
Обнаружение сарказма	347
Смысловое основание языка	347
Скимминг текста с помощью LSTM	348
Новые модели машинного обучения	348
Фазированные LSTM	349
Расширенные рекуррентные нейронные сети (DRNN)	350
Заключение	351
Литература	351

Приложение. Математические основы и углубленное

изучение TensorFlow	354
Основные структуры данных	354
Скаляр	354
Векторы	354
Матрицы	355
Индексы матрицы	355
Специальные типы матриц	356
Тождественная матрица	356
Диагональная матрица	356
Тензоры	357
Тензорные и матричные операции	357
Транспонирование	357
Умножение	358
Поэлементное умножение	358
Обратная матрица	359
Нахождение обратной матрицы – сингулярное разложение (SVD)	360
Нормы	360
Определитель	361

12 ❖ Содержание

Вероятность.....	361
Случайные величины	362
Дискретные случайные величины	362
Непрерывные случайные величины	362
Функция вероятности масса/плотность.....	362
Условная вероятность.....	364
Совместная вероятность	364
Предельная вероятность	365
Правило Байеса.....	365
Введение в Keras	365
Введение в библиотеку TensorFlow seq2seq.....	367
Определение вложений для кодера и декодера	367
Объявление кодера.....	368
Объявление декодера	369
Визуализация представлений слов с помощью TensorBoard	370
Первые шаги с TensorBoard.....	370
Сохранение представлений слов и визуализация в TensorBoard.....	371
Заключение	374
Предметный указатель	376