Отзывы о книге

Эта книга – идеальная отправная точка для практиков и разработчиков, которые хотят освоить языковую модель GPT-3 и научиться создавать приложения на API OpenAI.

– Питер Велиндер,
вице-президент по продукту и партнерским отношениям, OpenAI

Главная особенность этой книги в том, что ее могут прочитать люди с самым разным техническим образованием и создать решения мирового уровня с использованием ИИ.

– Ноа Гифт,
исполнительный директор Университета Дьюка,
основатель Pragmatic AI Labs

Если вы хотите использовать GPT-3 или любую другую большую языковую модель для создания своего приложения или службы, в этой книге найдется все, что вам нужно. В книге подробно рассматривается GPT-3, и примеры использования помогут вам применить эти знания к вашему продукту.

– Дэниел Эриксон,
основатель и генеральный директор Viable

Авторы проделали замечательную работу по изучению технических и социальных аспектов использования GPT-3. Прочитав эту книгу, вы будете уверенно рассуждать о современном состоянии искусственного интеллекта.

– Брэм Адамс,
основатель Steganography

Отличная книга для начинающих! В ней даже есть мемы и очень нужная глава об ИИ и этике, но ее главное достоинство – пошаговые процедуры работы с GPT-3.

– Рикардо Хосе Лима,
профессор лингвистики Университета Эстадо-ду, Рио-де-Жанейро

Это всестороннее глубокое погружение в работу с одной из ключевых генеративных моделей обработки естественного языка с практическим акцентом на том, как использовать API OpenAI и интегрировать его в ваши собственные приложения. Помимо очевидной технической ценности, я считаю особенно важными изложенные в последних главах мысли в отношении предубеждений и конфиденциальности моделей и их роли в демократизации ИИ.

– Рауль Рамос-Поллан,
профессор искусственного интеллекта
Университета Антиокии в Медельине, Колумбия

Содержание

От издательства.....11

Благодарности.....12

Об авторах.....14

Предисловие.....15

Глава 1. Революция большой языковой модели.....17

Что скрывается за кулисами NLP.....18
Языковые модели становятся больше и лучше.....20
Что скрывается за названием GPT-3?.....21
Генеративные модели.....21
Предварительно обученные модели.....22
Модели-трансформеры.....25
Модели для преобразования последовательности в последовательность.....25
Механизм внимания модели-трансформера.....27
GPT-3: краткая история.....28
GPT-1.....28
GPT-2.....29
GPT-3.....29
Доступ к API OpenAI.....33

Глава 2. Начало работы с API OpenAI.....37

Playground.....37
Особенности составления текстовых запросов.....41
Базовые модели.....52
Davinci.....53
Curie.....53
Babbage.....54
Ada.....54
Серия Instruct.....54

Конечные точки....56
List models (список моделей)....56
Retrieve model (получить модель)....57
Completions (завершения)....57
Files (файлы)....57
Embeddings (встраивания)....59
Настройка GPT-3....60
Примеры приложений на основе настраиваемых моделей GPT-3....61
Как настроить GPT-3 для вашего приложения....62
Подготовка и загрузка обучающих данных....62
Обучение новой настроенной модели....63
Использование точной модели....64
Токены....65
Расценки....67
Производительность GPT-3 в стандартных задачах NLP....69
Классификация текстов....70
Классификация без ознакомления....70
Классификация с однократным и ограниченным ознакомлением....71
Пакетная классификация....73
Распознавание именованных сущностей....74
Обобщение текста....75
Генерация текста....78
Генерация статьи для сайта....79
Генерация сообщений в социальных сетях....80
Заключение....80

Глава 3. GPT-3 и программирование....82

Как использовать API OpenAI с Python?....82
Как использовать API OpenAI с Go?....86
Как использовать API OpenAI с Java?....89
Sandbox GPT-3 на базе Streamlit....91
Заключение....94

Глава 4. GPT-3 как инструмент стартапов нового поколения....95

Модель как услуга....96
Стартапы нового поколения: примеры из практики....99

Творческие приложения GPT-3: Fable Studio 100
Приложения анализа данных GPT-3: Viable 105
Приложения чат-ботов GPT-3: Quickchat 107
Маркетинговые приложения GPT-3: Copysmith 111
Документирование приложений GPT-3: Stenography 113
Взгляд инвестора на экосистему стартапов вокруг GPT-3 116
Заключение 117

Глава 5. GPT-3 как новый этап корпоративных инноваций 119

Практический пример: GitHub Copilot 121
Как это работает 122
Разработка Copilot 124
Что означает программирование с малым кодом / без кода? 125
Масштабирование с помощью API 126
Каковы перспективы развития Github Copilot? 127
Практический пример: Algolia Answers 128
Оценка возможностей NLP 129
Конфиденциальность данных 130
Стоимость 130
Скорость и задержка 131
Первые уроки 132
Практический пример: Microsoft Azure OpenAI 133
Microsoft и OpenAI: предсказуемое партнерство 133
Собственный API OpenAI для Azure 134
Управление ресурсами 135
Безопасность и конфиденциальность данных 136
Модель как услуга на уровне предприятия 137
Другие службы искусственного интеллекта и машинного обучения Майкрософт 138
Совет для предприятий 139
OpenAI или служба Azure OpenAI: что следует использовать? 140
Заключение 141

Глава 6. GPT-3: хорошая, плохая, ужасная 142

Борьба с предвзятостью ИИ 143
Подходы к борьбе с предвзятостью 146
Некачественный контент и распространение дезинформации 150
Зеленый след LLM 159

Действуйте осторожно 161
Заключение 162

Глава 7. Демократизация доступа к искусственному интеллекту 164

Нет кода – нет проблем! 165
Доступ и модель как услуга 168
Заключение 169

Предметный указатель 171

Об авторах

Сандра Кублик – предприниматель в области ИИ, популяризатор и общественный деятель, которая продвигает бизнес-инновации, связанные с ИИ. Наставник и тренер нескольких компаний, занимающихся ИИ, соучредитель программы ИИ-акселераторов для стартапов и сообщества хакатонов ИИ Deep Learning Labs. Она является активным представителем сообщества NLP и генеративного ИИ. Ведет канал на YouTube, где берет интервью у различных действующих лиц экосистемы стартапов и обсуждает новаторские тенденции в области искусственного интеллекта с помощью забавного и образовательного контента.

Шубхам Сабу занимался разными видами деятельности, от специалиста по данным до консультанта по ИИ в известных фирмах по всему миру, где участвовал в разработке общеорганизационных стратегий работы с данными и технологической инфраструктуры для создания и масштабирования практики обработки данных и машинного обучения с нуля. Его работа в качестве популяризатора ИИ привела к появлению собственной широкой аудитории, где он продвигает идеи применения ИИ. Движимый страстью к изучению нового и обмену знаниями с сообществом, он ведет технические блоги о достижениях в области ИИ и экономических последствиях этого. В свободное время путешествует по стране, что позволяет ему погрузиться в разные культуры и развить свое мировоззрение на основе опыта.

Предисловие

Знаменитая GPT-3, или Generative Pretrained Transformer 3, представляет собой большую языковую модель на основе архитектуры Transformer, разработанную OpenAI. Она состоит из ошеломляющих 175 млрд параметров. Любой желающий может получить доступ к этой огромной языковой модели через API OpenAI – простой в использовании пользовательский интерфейс «текст на входе – текст на выходе» без каких-либо серьезных технических требований. Это первый случай в истории, когда модель искусственного интеллекта такого масштаба была размещена на удаленной платформе и доступна для широкой публики с помощью простого вызова API. Этот новый режим доступа называется «модель как услуга» (model-as-a-service, MaaS). Из-за этого невиданного ранее режима доступа многие люди, включая авторов этой книги, рассматривают GPT-3 как первый шаг к демократизации искусственного интеллекта (ИИ).

С появлением GPT-3 стало проще, чем когда-либо, создавать приложения ИИ. Эта книга в деталях покажет вам, как легко начать работу с API OpenAI. Кроме того, мы познакомим вас с инновационными способами использования этого инструмента в разных областях. Мы рассмотрим успешные стартапы, созданные на основе GPT-3, и корпорации, использующие его в своей продуктовой линейке, а также обсудим проблемы и перспективы развития.

Эта книга предназначена для людей с любым образованием и любого рода занятий, а не только для технических специалистов. Она будет особенно полезна, если вы:

- специалист по обработке данных, желающий приобрести навыки в области ИИ;
- предприниматель, который хочет построить следующий проект в области ИИ;
- руководитель компании, который хочет расширить свои знания об искусственном интеллекте и использовать их для принятия ключевых решений;
- писатель, подкастер, менеджер социальных сетей или другой создатель языковых продуктов, желающий использовать лингвистические возможности GPT-3 в творческих целях;

- любой, у кого есть идея, основанная на искусственном интеллекте, которая когда-то казалась технически невозможной или слишком дорогой для реализации.

Первая часть книги посвящена основам API OpenAI. Во второй части книги мы исследуем пеструю экосистему, органично и стремительно возникшую вокруг GPT-3.

В *главе 1* изложен контекст и основные определения, необходимые для комфортного изучения дальнейших тем. В *главе 2* мы глубоко погружаемся в API, разбивая его на наиболее важные элементы, такие как базовые модели и конечные точки, описывая их назначение и способы использования для читателей, которые хотят взаимодействовать с ними на более глубоком уровне. *Глава 3* содержит простой и интересный рецепт для вашего первого приложения на базе GPT-3.

Затем, переместив акцент на увлекательную экосистему ИИ, в *главе 4* мы берем интервью у создателей некоторых из самых успешных продуктов и приложений на основе GPT-3 и спрашиваем их о проблемах и опыте взаимодействия с моделью в коммерческом масштабе. В *главе 5* будет рассказано, как предприятия относятся к GPT-3 и каков потенциал внедрения этой модели. В *главе 6* мы обсуждаем потенциально проблематичные последствия более широкого внедрения GPT-3, такие как неправомерное использование и предвзятость, а также прогресс в решении этих проблем. Наконец, в *главе 7* мы заглядываем в будущее, знакомя вас с наиболее интересными тенденциями и возникающими возможностями, по мере того как GPT-3 все шире внедряется в коммерческую экосистему.

1

Революция большой языковой модели

«искусство – это обломки от столкновения души и мира»
«технологии стали мифом современного мира»
«революции начинаются с вопроса, но не заканчиваются ответом»
«природа украшает мир разнообразием»

Твиты, сгенерированные нейросетью GPT-3

Представьте, что вы проснулись прекрасным солнечным утром. Сегодня понедельник, и вы знаете, что неделя будет беспокойной. Ваша компания собирается запустить новое приложение для отслеживания личной продуктивности под названием Taskr и начинает кампанию в социальных сетях, чтобы рассказать миру о вашем гениальном продукте.

На этой неделе ваша главная задача – написать и опубликовать серию интересных постов в блоге. Вы начинаете с составления списка дел:

- написать информативную и забавную статью о лайфхаках для повышения производительности с упоминанием о Taskr. Не более 500 слов;
- создать список из пяти броских заголовков статей;
- выбрать визуальное оформление.

Вы нажимаете клавишу ввода, делаете глоток кофе и наблюдаете, как на вашем экране возникает статья, предложение за предло-

жением, абзац за абзацем. Через 30 секунд у вас готов содержательный высококачественный пост в блоге, идеальный старт для вашей серии публикаций в социальных сетях. Современное и красочное визуальное оформление привлекает внимание читателей. Готово! Вы выбираете лучшее название из пяти предложенных вариантов и приступаете к публикации.

Это не фантазия из далекого будущего, а зарисовка новой реальности, ставшей возможной благодаря достижениям в области искусственного интеллекта. Пока вы читаете эту книгу, одно за другим появляются новые приложения для креативной генерации текста и изображений, доступные всем желающим.

GPT-3 – это передовая языковая модель, созданная компанией OpenAI, которая находится на переднем крае исследований и разработок в области искусственного интеллекта. Первый официальный релиз OpenAI, в котором объявляется о создании GPT-3, был выпущен в мае 2020 года, а уже в июне 2020 года был открыт доступ к GPT-3 через API OpenAI. С момента запуска GPT-3 во всем мире были придуманы сотни, если не тысячи интересных применений этой модели в самых разных областях, включая технологии, искусство, литературу, маркетинг… и этот список постоянно растет.

GPT-3 может с невероятной легкостью решать общие языковые задачи, такие как создание и классификация текста, свободно перемещаясь между различными стилями текста и целями. Круг задач, которые она может решить, огромен.

В этой книге мы предлагаем вам подумать о том, какие задачи вы могли бы самостоятельно решить с помощью GPT-3. Мы обещаем рассказать вам, что это за модель и как ее использовать, но сначала хотим получше ввести вас в тему. В оставшейся части данной главы мы обсудим, откуда взялась эта технология, как она устроена, с какими задачами она лучше всего справляется и какие потенциальные риски она несет. Давайте пойдем короткой дорогой и начнем прямо с *обработки естественного языка* (natural language processing, NLP), посмотрим, как с ней связаны *большие языковые модели* (large language model, LLM) и GPT-3.

Что скрывается за кулисами NLP

NLP – это область информационных технологий, посвященная взаимодействию между компьютерами и человеческими языками. Цель исследователей – создать системы, способные эффективно

и качественно обрабатывать естественный язык, с помощью которого люди общаются друг с другом.

NLP сочетает в себе компьютерную лингвистику (моделирование человеческого языка на основе правил) с машинным обучением для создания интеллектуальных машин, способных определять контекст и понимать смысл естественного языка.

Машинное обучение – это ветвь ИИ, в которой исследователи развивают способность машин решать различные задачи с помощью опыта, без явного программирования. *Глубокое обучение* – это область машинного обучения, которая основана на использовании глубоких нейронных сетей, смоделированных по образцу человеческого мозга, для выполнения сложных задач с минимальным вмешательством человека.

Глубокое обучение появилось в 2010-х годах, и спустя некоторое время были созданы большие языковые модели на основе плотных нейронных сетей, состоящих из тысяч или даже миллионов простых рабочих элементов, называемых искусственными нейронами. Нейронные сети стали первым значительным прорывом в области NLP, позволив реализовать сложную обработку естественного языка, что до той поры считалось возможным только в теории. Второй важной вехой стало появление *предварительно обученных моделей* (таких как GPT-3), которые впоследствии можно точно настроить для различных задач, что позволяет сэкономить много часов обучения. (Предварительно обученные модели мы обсудим позже в этой главе.)

NLP лежит в основе многих прикладных применений ИИ, таких как:

Обнаружение спама
: Система фильтрации спама в вашем почтовом ящике использует NLP, чтобы определить, какие электронные письма выглядят подозрительно, и отправить их в корзину.

Машинный перевод
: Google Translate, DeepL и другие программы машинного перевода используют NLP для перевода предложений в почти произвольных языковых парах.

Виртуальные помощники и чат-боты
: В эту категорию попадают чат-боты наподобие Alexa, Siri, Google Assistant и многочисленные службы поддержки клиентов по всему миру. Они используют NLP, чтобы понимать и анализировать смысл обращения, определять приоритетность вопросов и запросов пользователей и быстро и правильно реагировать на них.

Анализ настроений в социальных сетях
: Маркетологи собирают в социальных сетях сообщения о конкретных брендах, темы разговоров и ключевые слова, а затем используют NLP для анализа индивидуального и коллективного отношения людей к бренду. Это помогает брендам в исследовании клиентов, оценке своего имиджа и определении социальной динамики.

Обобщение текста
: Обобщение текста – это уменьшение его размера при сохранении ключевой информации и основного смысла. Наиболее распространенными примерами обобщения текста являются заголовки новостей, анонсы фильмов, информационные бюллетени, финансовые обзоры, анализ юридических контрактов, сводки писем в электронной почте и приложения, доставляющие ленты новостей, отчеты и электронные письма.

Семантический поиск
: Семантический поиск использует глубокие нейронные сети для интеллектуального поиска данных. Вы взаимодействуете с ним каждый раз, когда выполняете поиск в Google. Семантический поиск полезен при поиске чего-либо на основе контекста, а не определенных ключевых слов.

«Мы взаимодействуем с другими людьми посредством языка», – говорит Янник Килчер (https://www.youtube.com/@YannicKilcher), один из самых популярных ютуберов и авторитетов в области NLP. «Язык является частью каждой бизнес-транзакции, любого совместного действия людей, и даже с машинами мы взаимодействуем посредством того или иного языка, будь то программа либо пользовательский интерфейс». Поэтому неудивительно, что компьютерная обработка естественного языка стала источником самых захватывающих открытий и местом самых впечатляющих применений ИИ за последнее десятилетие.

Языковые модели становятся больше и лучше

Моделирование языка – это задача присвоения вероятности последовательности слов в тексте на определенном языке. Основываясь на статистическом анализе существующих текстовых последова-

тельностей, простые языковые модели могут рассматривать слово и предсказывать следующее слово (или слова), которое, скорее всего, последует за ним. Чтобы создать языковую модель, которая успешно предсказывает последовательности слов, вы должны обучить ее на больших наборах данных.

Языковые модели – это жизненно важный компонент приложений для обработки естественного языка. Их можно рассматривать как инструмент статистического прогнозирования, получающий текст на входе и выдающий прогноз на выходе. Наверняка вы хорошо знакомы с этим инструментом в виде функции автозавершения в телефоне. Например, если вы напечатаете слово «добрый», автозавершение предложит варианты «человек», «день» и «путь».

До GPT-3 не существовало общей языковой модели, которая могла бы хорошо выполнять ряд задач NLP. Языковые модели были разработаны для выполнения *одной* конкретной задачи NLP, такой как генерация текста, обобщение или классификация. В этой книге мы обсудим экстраординарные возможности GPT-3 как общей языковой модели. Мы начнем эту главу с того, что познакомим вас с каждой буквой в аббревиатуре «GPT», чтобы показать, что они обозначают и из каких элементов построена знаменитая модель. Мы дадим краткий обзор истории и покажем, как и почему модели преобразования последовательностей, которые сегодня блистают в различных приложениях, достигли такого успеха. После этого мы расскажем вам о важности доступа к API и о том, как он развивался в зависимости от требований пользователей. Мы рекомендуем зарегистрировать учетную запись на сайте OpenAI, прежде чем переходить к остальным главам.

Что скрывается за названием GPT-3?

Название GPT-3 расшифровывается как «Generative Pre-trained Transformer 3» (генеративный предварительно обученный трансформер). Давайте рассмотрим все эти термины по порядку – это поможет нам понять принцип работы GPT-3.

Генеративные модели

GPT-3 – это *генеративная модель*, поскольку она генерирует текст. Генеративное моделирование – это раздел статистического моделирования. Это метод математической аппроксимации мира.

Нас окружает невероятное количество доступной информации – как в физическом, так и в цифровом мире. Сложность заключается в разработке интеллектуальных моделей и алгоритмов, способных анализировать и понимать эту сокровищницу данных. Генеративные модели являются одним из наиболее многообещающих подходов к достижению этой цели[1].

Чтобы обучить модель, вы должны подготовить и предварительно обработать *обучающий набор данных* – набор примеров, которые помогают модели научиться выполнять определенную работу. Обычно обучающий набор представляет собой большой объем данных в какой-то конкретной области: например, миллионы изображений автомобилей, чтобы научить модель распознавать автомобиль на незнакомых картинках. Обучающие данные могут принимать разнообразную форму. Это могут быть, например, текстовые предложения на естественном языке или фрагменты звуковых файлов (сэмплы). После того как вы показали модели множество примеров, она должна научиться генерировать аналогичные данные – в этом предназначение генеративной модели.

Предварительно обученные модели

Вы слышали о теории 10 000 часов? В своей книге «Выбросы: история успеха» Малкольм Гладуэлл утверждает, что отработки любого навыка в течение 10 000 часов достаточно, чтобы стать экспертом[2]. Это «экспертное» знание закрепляется в связях, которые ваш человеческий мозг развивает между своими нейронами. Модель ИИ делает нечто подобное.

Чтобы создать хорошо работающую модель, ее необходимо обучить с использованием определенного набора переменных, называемых *параметрами*. Процесс определения идеальных параметров для вашей модели называется *обучением*. Модель постепенно усваивает значения параметров, проходя через последовательные итерации обучения.

Глубокой модели, состоящей из множества нейронных слоев с миллионами нейронов, требуется много времени, чтобы найти эти идеальные параметры. Обучение – это длительный процесс, который в зависимости от задачи может длиться от нескольких

---

[1] Андрей Карпати (Andrej Karpathy) и др., публикация в блоге о генеративных моделях, источник: https://openИИ.com/blog/generative-models/.

[2] Malcolm Gladwell, *Outliers: The Story of Success* (Little, Brown, 2008).

часов до нескольких месяцев и требует огромных вычислительных мощностей. Очевидно, что нам очень пригодилась бы возможность повторно использовать результаты этого длительного процесса обучения для других задач. И здесь на помощь приходят предварительно обученные модели.

Если продолжить аналогию с теорией 10 000 часов Гладуэлла, то предварительно обученная модель – это базовый навык, который вы развиваете, чтобы легче было перейти к другому навыку. Например, овладение навыком решения математических задач поможет вам быстрее научиться решать инженерные задачи. Сначала модель обучают (вы или кто-то другой) для более общей задачи, а затем ее можно настроить для решения различных частных задач. Вместо того чтобы создавать совершенно новую модель для решения своей задачи, вы можете использовать предварительно обученную модель, которая уже в общих чертах владеет необходимыми «навыками». Предварительно обученную модель можно настроить в соответствии с вашими конкретными потребностями, предоставив дополнительное обучение с помощью специального набора данных. Этот подход намного быстрее и эффективнее и позволяет повысить производительность по сравнению с построением модели с нуля.

Размер набора данных, на которых обучают модель, во многом зависит от задачи, которую вы решаете, и от ваших возможностей собрать или приобрести необходимые данные. Модель GPT-3 обучена на текстовом корпусе из пяти наборов данных: Common Crawl, WebText2, Books1, Books2 и Wikipedia.

Common Crawl

Корпус Common Crawl содержит петабайты данных, включая необработанные данные веб-страниц, метаданные и текстовые данные, собранные за восемь лет сканирования веб-страниц. Исследователи OpenAI используют проверенную и отфильтрованную версию этого набора данных.

WebText2

WebText2 – это расширенная версия набора данных WebText, внутреннего корпуса OpenAI, созданного путем очистки веб-страниц особенно высокого качества. Чтобы гарантировать качество, авторы извлекли данные по всем исходящим ссылкам с Reddit, которые получили как минимум три кармы (индикатор того, что другие пользователи сочли ссылку интересной, познавательной или просто забавной). WebText содержит 40 ГБ текста, извлеченного из этих 45 млн ссылок и более 8 млн документов.

Books1 и Books2

Books1 и Books2 представляют собой два текстовых корпуса, которые содержат тексты десятков тысяч книг по различным предметам.

Wikipedia

Коллекция, включающая все англоязычные статьи из свободной онлайн-энциклопедии Wikipedia (https://en.wikipedia.org/wiki/МИИп_Page) на момент завершения сбора данных GPT-3 в 2019 году. Этот набор данных насчитывает примерно 5,8 млн статей на английском языке (https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia).

В общей сложности обучающий корпус содержит около триллиона слов.

GPT-3 может распознавать и генерировать тексты не только на английском языке. В табл. 1.1 показана первая десятка языков, наиболее широко представленных в обучающем наборе данных GPT-3 (https://github.com/OpenAI/gpt-3/blob/master/dataset_statistics/languages_by_document_count.csv).

Таблица 1.1. Десять наиболее широко представленных языков в наборе данных GPT-3

	Язык	Количество документов	Доля от общего кол-ва документов, %
1.	Английский	235 987 420	93,68882
2.	Немецкий	3 014 597	1,19682
3.	Французский	2 568 341	1,01965
4.	Португальский	1 608 428	0,63856
5.	Итальянский	1 456 350	0,57818
6.	Испанский	1 284 045	0,50978
7.	Голландский	934 788	0,37112
8.	Польский	632 959	0,25129
9.	Японский	619 582	0,24598
10.	Датский	396 477	0,15740

Разрыв между английским и остальными языками огромен. Английский язык занимает первое место с 93 % набора данных; немецкий язык, занимающий второе место, составляет всего 1 %, но даже этого достаточно для создания качественного текста на немецком языке с определенным стилем и решения других за-

дач. То же самое касается и других языков в списке (Русский язык находится на 16-м месте и составляет 0,11478 % обучающего набора. – *Прим. перев.*).

Поскольку GPT-3 предварительно обучена на обширном и разнообразном корпусе текстов, она может успешно выполнять удивительное количество разнообразных заданий в области NLP без предоставления пользователями каких-либо дополнительных данных.

Модели-трансформеры

Нейронные сети лежат в основе глубокого обучения, а их название и, во многом, структура позаимствованы у человеческого мозга. Они состоят из сети нейронов, которые работают вместе. Достижения в области нейронных сетей могут повысить производительность моделей ИИ в различных задачах, побуждая ученых в области ИИ постоянно разрабатывать новые архитектуры для этих сетей. Одним из таких достижений является модель-трансформер, которая обрабатывает всю последовательность текста сразу, а не по одному слову за раз, и обладает выдающейся способностью понимать взаимосвязь между этими словами. Это изобретение сильно повлияло на область обработки естественного языка.

Модели для преобразования последовательности в последовательность

Исследователи из Google и Университета Торонто представили модель-трансформер в статье 2017 года:

> *Мы предлагаем новую простую сетевую архитектуру Transformer, основанную исключительно на механизмах внимания и полностью исключающую рекуррентные и сверточные компоненты. Эксперименты с двумя задачами машинного перевода показали, что эти модели обеспечивают выдающиеся результаты, в то же время они легче распараллеливаются и требуют значительно меньше времени для обучения*[1].

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakon Uszkoreit, Llion Jones, Aidan Gomez, Lukasz Kaiser, and Illia Polosukhin, *Attention Is All You Need* (https://arxiv.org/abs/1706.03762), Advances in Neural Information Processing Systems 30 (2017).

В основе моделей-трансформеров лежит архитектура *преобразователя последовательности в последовательность* (sequence-to-sequence transformer, или коротко Seq2Seq). Такие модели особенно эффективны в задачах машинного перевода, которые представляют собой преобразование последовательности слов на одном языке в последовательность слов на другом языке. Google Translate начал использовать модель на основе Seq2Seq в 2016 году.

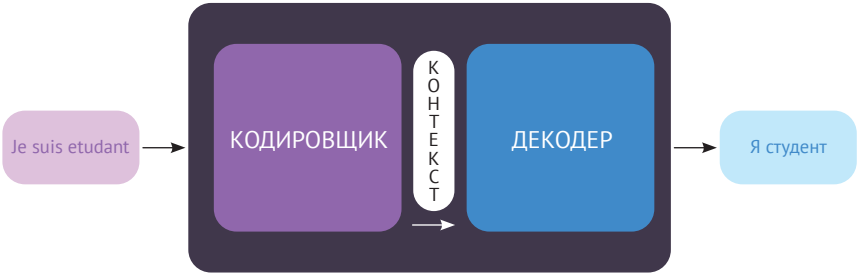

Рис. 1.1. Модель Seq2Seq (нейросетевой машинный перевод)[1]

Модели Seq2Seq состоят из двух компонентов: кодировщика и декодера. Кодировщик можно рассматривать как переводчика, для которого, к примеру, французский язык родной, а переводит он на корейский. Декодер – переводчик, для которого родным является русский язык, и он тоже владеет корейским. Для перевода с французского на русский кодировщик преобразует французское предложение в корейское (также известное как *контекст*, или *внутреннее представление*) и передает его декодеру. Поскольку декодер понимает корейский язык, он может преобразовать предложение с корейского на русский (т. е. восстановить из контекста). Таким образом, кодировщик и декодер вместе выполняют перевод с французского на русский[2], как показано на рис. 1.1. Причина использования промежуточного контекста в том, что это универсальное представление *смысла* любого языка, что позволяет строить произвольные языковые пары перевода на основе одного и того же представления.

---

1 Jay Alammar, *The Illustrated Transformer*, запись в блоге, источник: https://jalammar.github.io/illustrated-transformer/.

2 Jay Alammar, *The Illustrated Transformer*, запись в блоге, источник: https://jalammar.github.io/illustrated-transformer/.

Механизм внимания модели-трансформера

Архитектура Transformer была разработана для повышения качества ИИ в задачах машинного перевода. «Модели-трансформеры начинались как языковые модели, – объясняет Килчер, – и сперва они были небольшими, но потом выросли».

Чтобы эффективно использовать модели-трансформеры, крайне важно понять концепцию *внимания*. Механизм внимания имитирует концентрацию внимания человеческого мозга на определенных частях входной последовательности, используя вероятности для определения того, какие части последовательности наиболее важны на каждом этапе.

Например, возьмем предложение «Рыжая кошка села на коврик после того, как поймала мышь». Относится ли слово «рыжая» в этом предложении к «кошке» или к «мыши»? Модель-трансформер способна прочно связать слова «рыжая» и «кошка». Это и есть механизм внимания.

Важным моментом совместной работы кодировщика и декодера является тот факт, что кодировщик выделяет ключевые слова, связанные со значением предложения, и предоставляет их декодеру вместе с внутренним представлением. Эти ключевые слова облегчают декодеру понимание перевода, поскольку теперь он лучше различает значимые части предложения и термины, обеспечивающие контекст.

Модель-трансформер имеет два типа внимания: *самовнимание* (внутреннее внимание, связь слов внутри предложения) и *внимание кодировщика-декодера* (связь между словами из исходного предложения и словами из целевого предложения).

Механизм внимания помогает трансформеру отфильтровать шум и сосредоточиться на том, что имеет значение: суметь соединить два слова в семантической связи друг с другом, хотя эти слова не имеют очевидных маркеров, указывающих друг на друга.

Модели-трансформеры лучше работают в более крупных архитектурах и с большим объемом данных. Обучение на больших наборах данных и тонкая настройка под конкретные задачи заметно улучшают результаты. Модели-трансформеры лучше понимают контекст слов в предложении, чем любая другая нейронная сеть. GPT – это просто декодирующая часть трансформера.

Теперь, когда вы знаете, что означает «GPT», давайте поговорим о цифре 3 в названии, а также о цифрах 1 и 2.

GPT-3: краткая история

Модель GPT-3 была создана компанией OpenAI – пионером исследований в области искусственного интеллекта из Сан-Франциско. Заявленная миссия OpenAI состоит в том, чтобы «сделать так, чтобы искусственный интеллект приносил пользу всему человечеству». Заявленная миссия раскрывает понятие *искусственного общего интеллекта*: это тип ИИ, не ограничивающийся специализированными задачами и хорошо выполняющий множество разнообразных задач, как это делают люди.

GPT-1

Модель GPT-1 была представлена в июне 2018 года. Ключевой вывод разработчиков заключался в том, что сочетание архитектуры Transformer с предварительным обучением без учителя дало многообещающие результаты. Они писали, что модель GPT-1 была точно настроена для конкретных задач, чтобы добиться «глубокого понимания естественного языка».

GPT-1 стала важным шагом на пути к языковой модели с общими языковыми возможностями. Было доказано, что языковые модели поддаются эффективному предварительному обучению, что способствует хорошему *обобщению* (навыку работы с незнакомыми данными). Архитектура смогла выполнять различные задачи NLP лишь с небольшой тонкой настройкой.

В модели GPT-1 для обучения модели использовались набор данных BooksCorpus (https://yknzhu.wixsite.com/mbweb), который содержит около 7000 неопубликованных книг и механизм самовнимания в декодере преобразователя. Архитектура была аналогична оригинальному трансформеру со 117 млн параметров. Она проложила путь для будущих моделей с большими наборами данных и большим количеством параметров, позволяющими лучше использовать потенциал трансформерной архитектуры.

Одной из примечательных способностей GPT-1 была достойная производительность в задачах обработки естественного языка с нулевым обучением, таких как ответы на вопросы и анализ

настроений, достигнутая благодаря предварительному обучению. О*бучение без примеров*, или *обучение без ознакомления* (zero-shot learning), – это способность модели выполнять задачу, не будучи знакомой с другими примерами этой задачи. В случае нулевого обучения модели не предоставляются обучающие примеры, и она должна понимать текущую задачу на основе инструкций и нескольких примеров предыдущих задач.

GPT-2

В феврале 2019 года OpenAI представила модель GPT-2, которая больше, но в остальном очень похожа на GPT-1. Существенное отличие состоит в том, что GPT-2 может работать в многозадачном режиме. Было успешно доказано (https://cdn.OpenAI.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf), что языковая модель может хорошо выполнять несколько задач без доступа к обучающим примерам.

GPT-2 показала, что обучение на большем наборе данных и наличие большего количества параметров заметно улучшают способность языковой модели понимать контекст и позволяют превосходить другие модели даже в условиях отсутствия обучающих примеров для конкретной задачи. Стало окончательно ясно, что чем крупнее языковая модель, тем лучше она «понимает» естественный язык.

Чтобы создать обширный высококачественный набор обучающих данных, авторы просканировали Reddit и извлекли данные по исходящим ссылкам на статьи, за которые пользователи проголосовали на платформе. Получившийся набор данных WebText содержал 40 ГБ текстовых данных из более чем 8 млн документов, что намного больше, чем набор данных GPT-1. Модель GPT-2 был обучена на наборе данных WebText и содержала 1,5 млрд параметров, что в десять раз больше, чем у GPT-1.

GPT-2 оценивали по нескольким наборам задач, таких как понимание прочитанного, обобщение, перевод и ответы на вопросы.

GPT-3

Стремясь создать еще более надежную и мощную языковую модель, в OpenAI построили GPT-3. И набор данных, и сама модель примерно на два порядка больше, чем те, которые используются

для GPT-2: GPT-3 имеет 175 млрд параметров и обучена на комбинации пяти различных текстовых корпусов, что намного больше, чем набор данных, который применялся для обучения GPT-2. Архитектура GPT-3 во многом такая же, как GPT-2. Она хорошо работает с задачами NLP в сценариях без ознакомления (zero-shot) и с ознакомлением на нескольких примерах (few-shot).

GPT-3 умеет писать статьи, не отличимые от написанных человеком. Она также может выполнять на лету задачи, для которых не была специально обучена, например суммировать числа, составлять SQL-запросы и даже создавать код программ на языках React и JavaScript для задач, описанных на простом английском языке.

Примечание. Сценарии *few-shot*, *one-shot* и *zero-shot* – это особые случаи переноса знаний с ознакомлением. В режиме *few-shot* модель снабжена описанием задачи и таким количеством примеров, которое помещается в контекстное окно модели. Модель получает ровно один пример в режиме *one-shot* и ни одного примера в режиме *zero-shot*.

В своем заявлении о миссии компании OpenAI уделяет большое внимание доступности и этическим последствиям применения ИИ. Это видно по их решению сделать третью версию своей модели, GPT-3, доступной через открытый API. Благодаря доступу к API специальные программы-посредники облегчают связь между вебсайтом или приложением и пользователем.

Фактически API действует как средство связи между разработчиками и приложениями, позволяя им воплощать новые программные взаимодействия с пользователями. Доступ к GPT-3 через открытый API стал воистину революционным шагом. До 2020 года мощные модели ИИ, разработанные ведущими исследовательскими лабораториями, были доступны лишь немногим избранным исследователям и инженерам, работающим над этими проектами. API OpenAI предоставляет пользователям во всем мире невиданный ранее доступ к самой мощной в мире языковой модели через простой вход в систему. Бизнес-модель OpenAI заключается в создании новой парадигмы, которую они называют *model-as-a-service* (модель как сервис, MaaS), где разработчики могут платить за вызов API; мы рассмотрим этот вопрос более подробно в главе 3.

В ходе работы над созданием GPT-3 исследователи OpenAI экспериментировали с моделями разных размеров. Они взяли су-

ществующую архитектуру GPT-2 и увеличили количество параметров.

В результате этих экспериментов получилась модель с новыми и экстраординарными способностями. В то время как GPT-2 выполняла лишь некоторые задания с нулевыми примерами, GPT-3 может решать еще более сложные и незнакомые задачи, когда представлен пример контекста.

Исследователи OpenAI вполне обоснованно восхищаются тем, что простое увеличение параметров модели и размера обучающего набора привело к таким выдающимся достижениям. В целом они с оптимизмом надеются, что эти тенденции сохранятся даже для моделей, намного больших, чем GPT-3, что позволит создавать еще более сильные модели, способные к обучению на небольшом количестве примеров или совсем без примеров, просто путем точной настройки на выборке небольшого размера.

Пока вы читаете эту книгу, по оценкам экспертов (https://arxiv.org/abs/2101.03961), во всем мире создают и развертывают более триллиона параметрических языковых моделей. Мы вступили в золотой век больших языковых моделей, и пришло время вам стать его частью.

Примечание к переводу. Во время подготовки перевода этой книги был анонсирован запуск мультимодальной модели GPT-4, которая способна обрабатывать как текстовые, так и графические данные и выдавать ответ на естественном языке, а также в виде изображений и программного кода. Описание архитектуры, а также точное количество параметров модели на тот момент не были раскрыты.

Модель GPT-3 привлекла большое внимание общественности. В обзоре MIT Technology Review ее назвали одной из 10 революционных технологий 2021 года. Абсолютная гибкость GPT-3 в выполнении различных задач в сочетании с почти человеческой эффективностью и точностью производит ошеломляющее впечатление даже на искушенных пользователей. Как написал в Твиттере один из первых пользователей Аррам Сабети, «...это невероятно здорово» (рис. 1.2):

API OpenAI радикально изменил подход к NLP и привлек десятки тысяч бета-тестеров. За ними по пятам последовали инноваторы и стартапы, и многие комментаторы назвали GPT-3 «пятой промышленной революцией».

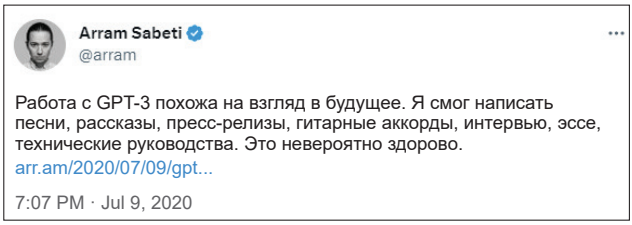

Рис. 1.2. Твит Аррама Сабети (https://twitter.com/arram/status/1281258647566217216?lang=en)

По данным OpenAI, всего за девять месяцев после запуска API пользователи построили с его помощью более трехсот предприятий. Хотя возникшая вокруг GPT-3 шумиха временами выглядит чрезмерной, многие эксперты уверены, что для нее есть все основания. Например, так думает Бакз Аван – разработчик, ставший предпринимателем и влиятельным лицом в сообществе разработчиков OpenAI API. У него есть канал на YouTube «Bakz T. Future» (https://www.youtube.com/user/bakztfuture). Аван утверждает, что, говоря про GPT-3 и другие модели, люди «недооценивают, насколько они на самом деле удобны, дружелюбны, приятны и мощны. Они почти шокируют».

Дэниел Эриксон, генеральный директор Viable, компании, предлагающей продукты на базе GPT-3, высоко оценивает способность модели извлекать информацию из больших наборов данных с помощью того, что он называет *разработкой на основе запросов* (prompt-based development):

> *Многие компании используют нейросетевую модель для сочинения рекламных текстов и контента веб-сайтов. Ключевая идея относительно проста: компания берет ваши данные, отправляет их модели в виде запроса и возвращает результат, полученный через API. Фактически она берет задачу, для решения которой достаточно одной подсказки API, оборачивает вокруг нее пользовательский интерфейс и доставляет пользователям.*

Проблема, которую Эриксон видит в этой области применения моделей, заключается в том, что она уже переполнена до краев, потому что привлекает множество амбициозных стартапов, конкурирующих

с одинаковыми услугами. Вместо этого Эриксон рекомендует рассмотреть другой вариант использования моделей, как это сделали в Viable. Способы применения моделей, основанные на данных, не так широко распространены, как получение быстрых подсказок, но они более прибыльны и позволяют оторваться от конкурентов.

Ключ, по словам Эриксона, заключается в том, чтобы создать большой и постоянно пополняемый набор данных, из которого GPT-3 извлекает ценную информацию. В этом заключается бизнес-модель Viable, позволившая им легко монетизировать услуги. «Люди платят гораздо больше за долгосрочные данные, чем за сиюминутные ответы», – объясняет Эриксон.

Технологические революции обычно влекут за собой новые противоречия и вызовы. GPT-3 – мощный инструмент в руках *любого*, кто пытается создать нарратив. Если не соблюдать осторожность и не придерживаться добрых намерений, мощная языковая модель может быстро превратиться в инструмент для проведения масштабных кампаний по дезинформации. Еще одной проблемой грозит стать массовое производство низкокачественного цифрового контента, который загрязняет информацию, доступную в интернете. Нельзя забывать и про разного рода предвзятость наборов данных, которую могут выявлять и усиливать новые технологии. Мы более подробно рассмотрим эти и другие проблемы в главе 6, а также обсудим усилия OpenAI по их устранению.

Доступ к API OpenAI

По состоянию на 2021 год на рынке уже было выпущено несколько фирменных моделей ИИ с бóльшим количеством параметров, чем у GPT-3, и их количество продолжает расти. Однако доступ к ним ограничен горсткой людей в стенах отделов исследований и разработок компаний или группой избранных тестировщиков, что делает невозможным оценку их эффективности в реальных задачах NLP.

Еще одним фактором доступности GPT-3 является его простой и интуитивно понятный пользовательский интерфейс, работающий по принципу банального ввода-вывода текста. Он не требует сложной тонкой настройки или знания механизма обновлений градиента, и вам не нужно быть экспертом, чтобы его использовать. Эта комбинация масштабируемых параметров и относительно открытого доступа делает GPT-3 самой интересной и, возможно, самой актуальной языковой моделью на сегодняшний день.

Из-за исключительных возможностей GPT-3 существуют значительные риски с точки зрения безопасности и неправильного использования, связанные с открытием исходного кода, которые мы рассмотрим в последней главе, – учитывая это, в OpenAI решили не публиковать исходный код GPT-3 и придумали уникальную, невиданную ранее модель совместного доступа через API.

Первоначально компания решила открыть доступ к API в виде ограниченного списка бета-тестеров. OpenAI открыла прием заявок на участие в тестировании, в котором люди должны были заполнить форму с подробным описанием своего опыта и причин запроса доступа к API. Только одобренным пользователям был предоставлен доступ к закрытой бета-версии API с интерфейсом под названием *Playground* (площадка для игр, песочница).

Буквально в первые дни список ожидания доступа к бета-версии GPT-3 достиг десятков тысяч человек. К чести OpenAI, они быстро отреагировали на такой наплыв желающих и начали добавлять пользователей в пакетном режиме. В OpenAI также внимательно следили за их действиями и отзывами о пользовательском интерфейсе API, чтобы постоянно улучшать его.

Благодаря прогрессу в мерах безопасности OpenAI отменила список ожидания в ноябре 2021 года. Для доступа к GPT-3 теперь достаточно простой регистрации. Это знаменательная веха в истории GPT-3, которую очень ждало сообщество. Чтобы получить доступ к API, перейдите на страницу регистрации (https://platform.OpenAI.com/signup), создайте бесплатную учетную запись и сразу же приступайте к экспериментам.

Примечание к переводу. Сайт и сервис OpenAI недоступны для пользователей с российским IP-адресом. Для создания учетной записи вам понадобится доступ через VPN с европейским IP и временный виртуальный телефонный номер в любой европейской стране, на который вы получите СМС с кодом подтверждения регистрации. В дальнейшем вам этот телефонный номер не понадобится, но доступ к API возможен только через VPN. Мы не будем здесь детально описывать использование виртуальных телефонных номеров, но отметим, что российским пользователям доступно множество сервисов, предоставляющих виртуальные телефонные номера для получения СМС в различных странах с оплатой российскими банковскими картами, включая «Мир», по очень доступной цене. Как показал опыт, для регистрации и входа на сайт OpenAI можно использовать российский аккаунт Google. Процедура регистрации на сайте OpenAI детально описана в различных блогах российских авторов.

Новые пользователи получают начальный кредит, что позволяет им свободно экспериментировать с API. Кредит эквивалентен созданию текстового контента величиной приблизительно в три романа средней длины.

Примечание к переводу. По состоянию на март 2023 г. новым пользователям на счет зачислялся грант в размере 18 долларов, действующий в течение 40 дней, после чего неиспользованный остаток гранта «сгорает». Имейте это в виду при проведении экспериментов за счет гранта. Разумеется, условия предоставления гранта могут в любой момент измениться без предупреждения со стороны OpenAI.

После того как бесплатные кредиты исчерпаны, пользователи начинают платить за использование или, если у них есть уважительная причина, они могут запросить дополнительные кредиты в службе поддержки клиентов OpenAI API.

Примечание к переводу. К сожалению, на данный момент для российских пользователей не существует простого способа оплаты сервисов OpenAI после исчерпания гранта. Необходимо иметь банковскую карту Visa или Mastercard, выпущенную западным банком, поэтому проблему оплаты каждый пользователь должен решать самостоятельно. Впрочем, начального гранта вполне достаточно для экспериментов и знакомства с API OpenAI.

OpenAI стремится обеспечить ответственное создание приложений на базе API. По этой причине компания предоставляет инструменты (https://platform.OpenAI.com/docs/guides/moderation), примеры использования (https://platform.OpenAI.com/docs/guides/safety-best-practices) и руководства по использованию (https://platform.OpenAI.com/docs/usage-policies), которые помогут разработчикам быстро и безопасно запускать свои приложения в производство.

Компания также создала руководство по контенту (https://platform.OpenAI.com/docs/usage-policies/content-guidelines), чтобы уточнить, для создания какого контента можно использовать OpenAI API. Чтобы помочь разработчикам убедиться, что их приложения используются по назначению, предотвратить возможное неправильное применение и соблюдать рекомендации по содержанию,

OpenAI предлагает бесплатный фильтр содержимого. Политика OpenAI запрещает использование API способами, которые не соответствуют принципам, описанным в уставе (https://OpenAI.com/charter/), включая создание контента, пропагандирующего ненависть, насилие или членовредительство или для преследования личностей по любым мотивам, влияния на политические процессы, распространения дезинформации, спам-контента и т. д.

После регистрации на сайте OpenAI вы можете перейти к главе 2, где мы обсудим различные компоненты API, интерфейс «песочницы» Playground и способы использования API для различных целей.