Посвящается Маттие, Джакомо и Микеле

«Счастье реально только тогда, когда есть с кем разделить его».
– Кристофер МакКэндлесс

Содержание

Предисловие

Не просто дать количественную оценку того, насколько обыденными стали такие термины, как *нейронные сети* и *глубокое обучение*, и, более конкретно, как эти технологии влияют на нашу жизнь. От автоматизации рутинных заданий до тиражирования трудных решений, помощи в управлении автомобилем (и людьми) до места назначения – мощь нейронных сетей и глубокого обучения в качестве методов, которые произведут революцию в области вычислений, находится лишь в стадии зарождения.

Вот почему эта книга так важна. Нейронные сети, искусственный интеллект (ИИ) и глубокое обучение не только автоматизируют рутинные задания и решения, облегчая их. Они также облегчают и поиск. Прежде состояние дел в области поиска информации использовало сложную линейную алгебру, включая в себя матричное умножение для обозначения сопоставления пользовательских запросов с документами. Сегодня вместо использования алгебраических и линейных моделей применяются, например, нейронные сети для распознавания сходства слов между документами после изучения способов суммирования документов в слова с использованием обособленных сетей. И это только одна область в процессе поиска, где используются ИИ и глубокое обучение.

В своей книге Томмазо Теофили использует практический подход, чтобы показать вам современное состояние использования нейронных сетей, искусственного интеллекта и глубокого обучения в разработке поисковых систем. Книга изобилует примерами и знакомит читателя с архитектурой современных поисковых систем, а также дает вам достаточно информации, чтобы понять, как и где подходит глубокое обучение и как это улучшает поиск. Автор показывает вам, где искусственный интеллект и глубокое обучение могут перегрузить ваш код и возможность поиска, начиная от создания вашей первой сети для поиска похожих слов в расширении запроса и заканчивая изучением векторного представления слов для поискового ранжирования, а также мультиязычного поиска и поиска по изображениям.

Эта книга написана настоящим первопроходцем в области открытого исходного кода. Томмазо – бывший председатель проекта Apache Lucene – механизма индексации поиска де-факто, который поддерживает Elasticsearch и Apache Solr. Он также внес большой вклад в понимание языков и перевод при разработке библиотеки Apache OpenNLP. Совсем недавно его кандидатура была предложена на пост председателя (инкубационного) проекта Apache Joshua для статистического машинного перевода.

Я знаю, что вы многому научитесь из этой книги, и я рекомендую ее, чтобы вы могли найти золотую середину между здравым смыслом, толкованиями теории сложности и реальным кодом, с которым вы можете экспериментировать, используя новейшие технологии глубокого обучения и поиска.

Наслаждайтесь. Я знаю, о чем говорю, потому что именно это я и делал!

Крис Мэттманн,
заместитель директора по технологиям и инновациям,
Лаборатория реактивного движения НАСА

От автора

Обработка естественного языка околдовала меня, как только я узнал о ней, почти 10 лет назад, когда учился в магистратуре. Обещание, что компьютеры могут помочь нам понять (уже даже тогда) огромное количество существующих текстовых документов, было похоже на волшебство. Я до сих пор помню, как было здорово видеть, как мои первые программы для ОЕЯ извлекают пусть даже смутно правильную и полезную информацию из нескольких текстовых документов.

Примерно в то же время на работе меня попросили проконсультировать клиента по поводу новой архитектуры поиска с открытым исходным кодом. Мой коллега, который был экспертом в этой области, был занят другим проектом, поэтому мне дали копию книги *Lucene в действии*[1], которую я изучал на протяжении пары недель. Затем меня отправили на работу консультантом. Спустя пару лет после того, как я проработал над проектом на базе Lucene/Solr, заработала новая поисковая система (и, насколько я знаю, она используется до сих пор). Не могу сказать, сколько раз нужно было настраивать алгоритмы поисковой системы из-за того или иного запроса или того или иного фрагмента проиндексированного текста, но мы заставили ее работать. Я мог видеть запросы пользователей и мог видеть данные, которые должны были быть извлечены, но минимальная разница в написании или пропуске определенного слова могла привести к тому, что очень релевантная информация не была бы отображена в результатах поиска. Поэтому, хотя я очень гордился своей работой, я продолжал задаваться вопросом, как сделать все возможное, чтобы избежать множества ручных вмешательств, которые менеджеры по программному продукту просили меня выполнить, чтобы обеспечить наилучший опыт взаимодействия.

Сразу же после этого я совершенно случайно оказался вовлеченным в машинное обучение благодаря первому онлайн-занятию по машинному обучению от Эндрю Ына (которое было основано на серии Coursera MOOC). Я был настолько очарован концепциями нейронных сетей, показанными на уроке, что решил самостоятельно попробовать реализовать небольшую библиотеку для нейронных сетей на Java, просто ради удовольствия (**http://svn.apache.org/repos/asf/labs/yay/**). Я начал искать другие онлайн-курсы, такие как курс Андрея Карпати по сверточным нейронным сетям для визуального распознавания и курс Ричарда Сохера по глубоким нейронным сетям для обработки естественного языка. С тех пор я продолжаю работать над поисковыми системами, обработкой естественного языка и глубоким обучением, в основном в открытом исходном коде.

Пару лет назад (!) издательство Manning обратилось ко мне с рецензией на книгу об ОЕЯ, и я был достаточно наивен, чтобы написать в нижней части своего обзора, что мне было бы интересно написать книгу о поисковых системах и нейронных сетях. Когда издательство снова обратилось ко мне, проявив интерес, я был немного удивлен и спросил себя, *действительно* ли я хочу написать книгу об этом? Я понял, что да, мне это было интересно.

---

[1] **http://www.manning.com/books/lucene-in-action-second-edition**.

Несмотря на то что глубокое обучение произвело революцию в компьютерном зрении и обработке естественного языка, многое еще предстоит обнаружить, когда речь идет о приложениях, использующихся в поиске. Я уверен, что мы не можем (пока еще?) полагаться на глубокое обучение, чтобы автоматически настраивать поисковые системы от своего имени, но это может помочь сделать работу пользователя поисковой системы более гладкой. Благодаря глубокому обучению мы можем делать в поисковых системах то, чего пока не можем делать с помощью других существующих методов, и можем использовать глубокое обучение, чтобы улучшить техники, которые мы уже используем в поисковых системах. Путь к тому, чтобы сделать поисковые машины более эффективными с помощью глубоких нейронных сетей, только начался. Надеюсь, вам понравится.

Благодарности

Прежде всего я хотел бы поблагодарить свою любимую жену Микелу за помощь и поддержку на протяжении всего этого долгого путешествия: спасибо за любовь, энергию и преданность делу в течение долгих дней, ночей и выходных, в ходе которых я писал эту книгу!

Спасибо Джакомо и Маттиа за то, что они помогли мне выбрать самую классную иллюстрацию для обложки, а также за те игры и смех, которые сопровождали меня, пока я пытался писать.

Я хотел бы поблагодарить своего отца за то, что он гордится мной, и его веру в меня.

Большое спасибо моему другу Федерико за его неустанные усилия по просмотру всех материалов (книга, код, изображения и т. д.) и за приятные обсуждения и обмен идеями. Огромное спасибо моим друзьям и коллегам Антонио, Франческо и Симоне за их поддержку, смех и советы. Также адресую благодарность своим товарищам по проекту Apache OpenNLP (**http://opennlp.apache.org**), Сунилу, Йорну и Кодзи, за то, что предоставили отзывы, советы и идеи, которые помогли придать очертания этой книге.

Я благодарю Криса Мэтманна за написание такого вдохновляющего пролога.

Я также благодарю Фрэнсис Лефковиц, моего редактора по разработке, за ее терпение и руководство на протяжении всего процесса написания книги, включая наши дискуссии о Стефе, К.Д. и Воинах. И я благодарю других людей из издательства Manning, благодаря которым стало возможным появление этой книги, включая издателя Марьян Бэйс и лиц из редакционной и производственной команд, которые работали за кадром. Кроме того, я благодарю технических рецензентов во главе с Иваном Мартиновичем – Абхинава Упадхая, Эля Кринкера, Альберто Симоеса, Альваро Фалькину, Эндрю Уилли, Антонио Магнаги, Криса Моргана, Джулиано Бертоти, Грега Занотти, Йеруна Бенкхуйзена, Крифа Дэвида, Люка Мартина Бира, Майкла Уолла, Михала Пашкевича, Мирко Кемпфа, Паули Сутелайнен, Симона Русо, Срдана Дукича и Урсина Стаусса – и участников форума. Что касается технической стороны, выражаю благодарность Михилу Тримпе, который был техническим редактором книги, и Карстену Стрёбеку, который работал техническим корректором книги.

Наконец, я хотел бы поблагодарить сообщества Apache Lucene и Deeplearning4j за то, что предоставили такие превосходные инструменты, и за дружескую поддержку пользователей.

Об этой книге

Создание поисковых систем с использованием методов глубокого обучения – это практическая книга о том, как использовать (глубокие) нейронные сети для создания эффективных поисковых систем. В ней рассматривается несколько компонентов поисковой системы, дается представление о том, как они работают, и рекомендации о том, как можно использовать нейронные сети в каждом контексте. Основное внимание уделяется практическому объяснению методов поиска и глубокого обучения на базе примеров. Большинство из них сопровождается кодом. В то же время, когда это необходимо, приводятся ссылки на соответствующие исследовательские работы, чтобы побудить вас читать больше и углублять свои знания по конкретным темам. Нейронные сети и темы, связанные с поиском, объясняются на протяжении всей книги.

Прочитав ее, вы получите четкое представление об основных проблемах, связанных с поисковыми системами, о том, как они обычно решаются, и о том, как в этом может помочь глубокое изучение. Вы получите четкое представление о ряде различных методов глубокого обучения и о том, где они вписываются в контекст поиска. Вы также познакомитесь с библиотеками Lucene и Deeplearning4j. Кроме того, вы выработаете практическое отношение к тестированию эффективности нейронных сетей (вместо того чтобы рассматривать их как волшебство) и измерению их затрат и выгод.

Для кого предназначена эта книга

Данная книга предназначена для читателей, владеющих программированием на среднем уровне. Еще лучше, если вы хорошо разбираетесь в программировании на языке Java, интересуясь или активно участвуя в разработке поисковых систем. Вам следует прочитать эту книгу, если вы хотите сделать свою поисковую систему более эффективной, чтобы предоставлять релевантные результаты и, следовательно, сделать ее более полезной для конечных пользователей.

Даже если у вас нет такого опыта, вы будете знакомиться с основными понятиями, касающимися поисковых систем, на протяжении всей книги, когда будет затрагиваться каждый конкретный аспект поиска. Аналогично вам не обязательно иметь познания в области машинного или глубокого обучения. В этой книге будут представлены все необходимые основы машинного и глубокого обучения, а также практические советы, касающиеся применения глубокого обучения в поисковых системах в случаях реальной эксплуатации.

Вы должны быть готовы взять в руки код и расширить существующие библиотеки с открытым исходным кодом для реализации алгоритмов глубокого обучения для решения задач поиска.

Дорожная карта

Эта книга состоит из трех частей:

- первая часть знакомит вас с основными понятиями поиска, машинного и глубокого обучения. В главе 1 рассказывается об обосновании примене-

ния методов глубокого обучения для поиска проблем, затрагивая проблемы в отношении наиболее распространенных подходов к поиску информации. В главе 2 приводится первый пример того, как использовать модель нейронной сети для повышения эффективности поисковой системы путем генерации синонимов из данных;

- вторая часть посвящена общим задачам поисковых систем, которые можно лучше решать с помощью глубоких нейронных сетей. Глава 3 знакомит вас с использованием рекуррентных нейронных сетей для генерации запросов, альтернативных тем, которые вводят пользователи. Глава 4 посвящена задаче предоставления других предложений, в то время как пользователь набирает запрос, с помощью глубоких нейронных сетей. Глава 5 рассказывает о моделях ранжирования, в частности о том, как предоставить более релевантные результаты поиска, используя векторные представления слов. Глава 6 посвящена использованию векторных представлений документов как в функциях ранжирования, так и в контексте метода фильтрации на основе содержания, используемого при построении рекомендательных систем;
- в третьей части рассматриваются более сложные сценарии, такие как машинный перевод с использованием глубокого обучения и поиск изображений. В главе 7 рассказывается о мультиязычных возможностях поисковой системы с помощью подходов на базе нейронных сетей. Глава 8 посвящена поиску коллекции изображений на основе их содержимого, основанной на моделях глубокого обучения. В главе 9 обсуждаются темы, связанные с реальной эксплуатацией, такие как точная настройка моделей глубокого обучения и работа с постоянно поступающими потоками данных.

Сложность рассматриваемых тем и концепций возрастает в ходе прочтения книги. Если вы новичок в области глубокого обучения, поиска или того и другого, я настоятельно рекомендую сначала прочитать главы 1 и 2. В противном случае не стесняйтесь перепрыгивать и выбирать главы, основываясь на своих потребностях и интересах.

О КОДЕ

В этой книге предпочтение отдается фрагментам кода, а не подробным листингам, чтобы читатель мог быстро и легко понять, что и как делает код. Полный исходный код можно найти на странице книги на сайте издательства Manning: **www.manning.com/books/deep-learning-for-search**. Программное обеспечение также будет обновляться на официальной странице книги на GitHub (**https://github.com/dl4s**), включая исходный код на Java в книге (с использованием Apache Lucene и Deeplearning4j: **https://github.com/dl4s/dl4s**) и версию на Python для тех же алгоритмов (**https://github.com/dl4s/pydl4s**).

В примерах кода используется язык программирования Java и две библиотеки с открытым исходным кодом (по лицензии Apache): Apache Lucene (**http://lucene.apache.org**) и Deeplearning4j (**http://deeplearning4j.org**). Lucene является одной из наиболее широко используемых библиотек для создания поисковых систем, а Deeplearning4j на момент написания этих строк является лучшим выбором для нативной библиотеки Java для глубокого изучения. Вместе они позволят вам лег-

ко, быстро и без проблем проводить тестирование и экспериментировать с поиском и глубоким обучением.

Кроме того, многие специалисты, работающие над проектами, связанными с глубоким обучением, в настоящее время используют Python (с такими фреймворками, как TensorFlow, Keras, PyTorch и т. д.). Поэтому также предоставляется репозиторий Python, на котором размещены версии алгоритмов, подробно описанных в книге, на TensorFlow (**https://tensorflow.org**).

Исходный код в книге отформатирован `шрифтом фиксированной ширины, как этот`, чтобы отделить его от обычного текста. Во многих случаях первоначальный исходный код был переформатирован; мы добавили разрывы строк и переработали отступы, чтобы разместить доступное пространство страницы в книге. В редких случаях даже этого было недостаточно, и списки содержат маркеры продолжения строки (➥). Кроме того, комментарии в исходном коде часто удалялись из списков, когда описание кода приводится в тексте. Аннотации к коду сопровождают множество листингов, выделяя важные понятия.

Форум liveBook

Приобретение этой книги включает в себя бесплатный доступ к частному веб-форуму, организованному издательством Manning Publications, где вы можете оставлять комментарии о книге, задавать технические вопросы и получать помощь от автора и других пользователей. Чтобы получить доступ к форуму, перейдите по ссылке **https://livebook.manning.com/#!/book/deep-learning-for-search/discussion**. Вы можете узнать больше о форумах издательства и правилах поведения на странице **https://livebook.manning.com/#!/discussion**.

Обязательство Manning по отношению к нашим читателям состоит в том, чтобы обеспечить место, где может иметь место содержательный диалог между отдельными читателями и между читателями и автором. Это не обязательство какого-либо конкретного количества участия со стороны автора, чей вклад в форум остается добровольным (и неоплачиваемым). Мы предлагаем вам задавать автору сложные вопросы, чтобы его интерес не угас! Форум и архив предыдущих обсуждений будут доступны на сайте издателя, пока книга находится в печати.

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 прямо на нашем сайте **www.dmkpress.com**, зайдя на страницу книги, и оставить комментарий в разделе «Отзывы и рецензии». Также можно послать письмо главному редактору по адресу **dmkpress@gmail.com**, при этом напишите название книги в теме письма.

Если есть тема, в которой вы квалифицированы, и вы 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**.

Скачивание исходного кода примеров

Скачать файлы с дополнительной информацией для книг издательства «ДМК Пресс» можно на сайте **www.dmkpress.com** или **www.дмк.рф** на странице с описанием соответствующей книги.

Список опечаток

Хотя мы приняли все возможные меры для того, чтобы удостовериться в качестве наших текстов, ошибки все равно случаются. Если вы найдете ошибку в одной из наших книг – возможно, ошибку в тексте или в коде, – мы будем очень благодарны, если вы сообщите нам о ней. Сделав это, вы избавите других читателей от расстройств и поможете нам улучшить последующие версии этой книги.

Если вы найдете какие-либо ошибки в коде, пожалуйста, сообщите о них главному редактору по адресу **dmkpress@gmail.com**, и мы исправим это в следующих тиражах.

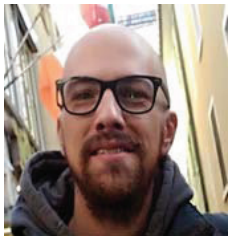
Нарушение авторских прав

Пиратство в интернете по-прежнему остается насущной проблемой. Издательства «ДМК Пресс» и Manning очень серьезно относятся к вопросам защиты авторских прав и лицензирования. Если вы столкнетесь в интернете с незаконно выполненной копией любой нашей книги, пожалуйста, сообщите нам адрес копии или веб-сайта, чтобы мы могли применить санкции.

Пожалуйста, свяжитесь с нами по адресу электронной почты **dmkpress@gmail.com** со ссылкой на подозрительные материалы.

Мы высоко ценим любую помощь по защите наших авторов, помогающую нам предоставлять вам качественные материалы.

Об авторе



Томмазо Теофили – инженер-программист со страстью к открытому исходному коду и машинному обучению. Будучи участником Apache Software Foundation, он участвует в ряде проектов с открытым исходным кодом, начиная от таких тем, как поиск информации (например, Lucene и Solr), и заканчивая обработкой естественного языка и машинным переводом (включая OpenNLP, Joshua и UIMA).

В настоящее время он работает в компании Adobe, разрабатывая компоненты инфраструктуры поиска и индексации, а также исследует области обработки естественного языка, поиска информации и глубокого изучения. Он выступал с докладами о поиске и машинном обучении на конференциях, включая BerlinBuzzwords, международную конференцию International Conference on Computational Science, ApacheCon, EclipseCon и др. Вы можете найти его в Twitter (**@tteofili**).

Об иллюстрации на обложке

Рисунок на обложке книги носит название «Одеяние китайской дамы». Иллюстрация взята из книги «Коллекция платьев разных народов, древних и современных» Томаса Джефферис (четыре тома), опубликованной в Лондоне между 1757 и 1772 годом. Титульный лист гласит, что это медные гравюры ручной работы, украшенные гуммиарабиком.

Томаса Джефферис (1719–1771) называли «географом короля Георга III». Он был английским картографом, который был ведущим создателем карт своего времени. Он гравировал и печатал карты для правительственных и других государственных учреждений и выпускал обширный спектр коммерческих карт и атласов, особенно касающихся Северной Америки. Его работа в качестве картографа пробудила интерес к местному дресс-коду, блистательно представленному в этой коллекции. Он был принят в тех землях, которые он исследовал и наносил на карту. Увлечение далекими землями и путешествия ради удовольствия были относительно новым явлением в конце XVIII века, и такие коллекции, как эта, были популярны, знакомя как туристов, так и путешественников, сидящих в креслах, с жителями других стран.

Разнообразие рисунков в этом издании Джефферис ярко свидетельствует об уникальности и индивидуальности народов мира около 200 лет назад. С тех пор дресс-код изменился, а богатое в ту пору разнообразие, в зависимости от региона и страны, исчезло. Сейчас часто трудно отличить жителей одного континента от другого. Возможно, пытаясь взглянуть на это с оптимизмом, мы обменяли культурное и визуальное разнообразие на более разнообразную личную жизнь – или на более разнообразную и интересную интеллектуальную и техническую жизнь.

В то время когда трудно отличить одну компьютерную книгу от другой, издательство Manning празднует изобретательность и инициативу компьютерного бизнеса с помощью обложек книг, основанных на богатом разнообразии жизни регионов двухвековой давности, которое ожило благодаря рисункам Джефферис.

Часть I

ПОИСК ВСТРЕЧАЕТСЯ С ГЛУБОКИМ ОБУЧЕНИЕМ

Настройка поисковых систем для эффективного реагирования на потребности пользователей – непростая задача. Традиционно многие внутренние настройки и корректировки, сделанные вручную, приходилось вносить в поисковую систему, чтобы она прилично работала при реальном сборе данных. С другой стороны, глубокие нейронные сети очень хорошо подходят для изучения полезной информации об огромных объемах данных. В этой первой части книги мы начнем изучать, как можно использовать поисковую систему в сочетании с нейронной сетью, чтобы обойти некоторые общие ограничения и предоставить пользователям более совершенные возможности поиска.

Глава 1

Поиск на основе нейронных сетей

О чем идет речь в этой главе:

- деликатное введение в основы поиска;
- важные проблемы в поиске;
- почему нейронные сети могут помочь поисковым системам быть более эффективными.

Предположим, вам нужно узнать что-то о последних научных открытиях в области искусственного интеллекта. Что вы будете делать, чтобы найти информацию? Сколько времени и сил требуется, чтобы получить факты, которые вы ищете? Если вы находитесь в (огромной) библиотеке, можно спросить библиотекаря, какие книги есть по этой теме, и он, вероятно, укажет на несколько книг, о которых он знает. В идеале библиотекарь подскажет определенные главы, которые нужно искать.

Звучит довольно просто. Но библиотекарь обычно происходит из другого контекста, в отличие от вас, то есть у вас и у библиотекаря могут быть разные мнения относительно того, что является важным. В библиотеке могут быть книги на разных языках, или библиотекарь может говорить на другом языке. Информация по этой теме может быть устаревшей, учитывая, что *последняя* является довольно относительным моментом во времени, и вы не знаете, когда библиотекарь в последний раз читал что-либо об искусственном интеллекте, или регулярно ли библиотека получает публикации в этой области. Кроме того, библиотекарь может не понять ваш запрос должным образом и подумать, что вы говорите об интеллекте с точки зрения психологии[1]. Пройдет несколько повторных циклов, прежде чем вы поймете друг друга и получите нужные вам фрагменты информации.

Затем, после всего этого вы можете обнаружить, что в библиотеке нет нужной вам книги; или информация может содержаться в нескольких книгах, и вы должны прочитать их все. Как это утомительно!

Только если вы сами не библиотекарь, именно это часто происходит в наши дни, когда вы что-то ищете в интернете. Хотя мы можем рассматривать интернет как единую огромную библиотеку, существует множество разных библиотекарей,

[1] Такое случилось со мной на самом деле.

которые помогут вам найти необходимую информацию: поисковые системы. Некоторые из них являются экспертами в определенных темах; другие знают только подмножество библиотеки или лишь одну книгу.

А теперь представьте, что некто, назовем его Робби, который уже знает о библиотеке и ее посетителях, может помочь вам обмениваться данными с библиотекарем, чтобы найти то, что вы ищете. Это поможет вам быстрее получить ответы. Робби может помочь библиотекарю понять запрос посетителя, например предоставив дополнительный контекст. Робби знает, о чем обычно читает посетитель, поэтому он пропускает все книги по психологии. Также, прочитав множество книг в библиотеке, Робби лучше понимает, что является важным в области искусственного интеллекта. Было бы чрезвычайно полезно иметь таких советников, как Робби, чтобы помочь поисковым системам работать лучше и быстрее, а также помогать пользователям получать больше полезной информации.

Эта книга посвящена использованию методов из области машинного обучения, называемой *глубоким обучением*, чтобы создавать модели и алгоритмы, которые могут влиять на поведение поисковых систем, чтобы сделать их более эффективными. Алгоритмы глубокого обучения будут играть роль Робби, помогая поисковой системе обеспечить лучший опыт поиска и предоставлять более точные ответы конечным пользователям.

Важно отметить, что глубокое обучение и *искусственный интеллект* – не одно и то же. Как видно по рис. 1.1, искусственный интеллект представляет собой огромную область для исследований. Машинное обучение является лишь частью этого, а глубокое обучение, в свою очередь, является подразделом машинного обучения. В основном глубокое обучение изучает, как заставить машины «учиться», используя модель вычислений глубоких нейронных сетей.

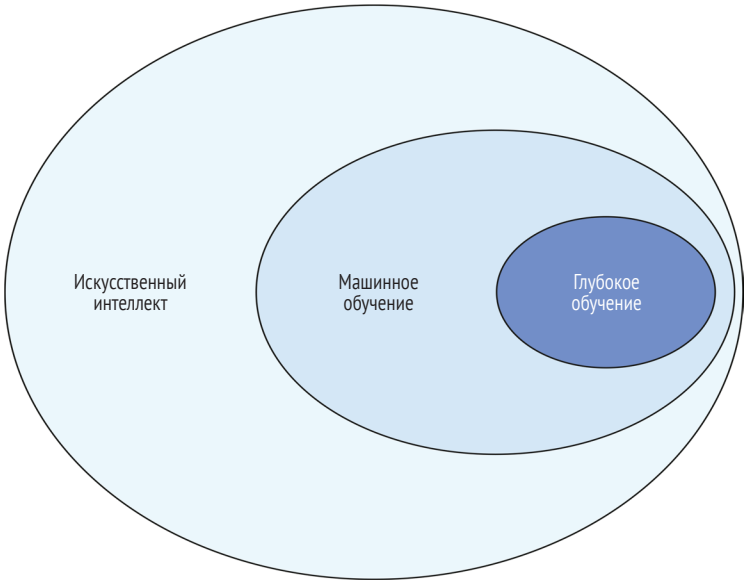


Рис. 1.1 ❖ Искусственный интеллект, машинное обучение и глубокое обучение

1.1. Нейронные сети и глубокое обучение

Цель этой книги – дать вам возможность использовать глубокое обучение в контексте поисковых систем, чтобы улучшить процесс поиска и его результаты. Даже если вы не собираетесь создавать еще одну поисковую систему в Google, вы должны научиться пользоваться методами глубокого обучения в небольших или средних поисковых системах, чтобы помочь пользователям в процессе поиска. Поиск на основе нейронных сетей должен помочь вам автоматизировать работу, которую в противном случае вам пришлось бы выполнять вручную. Например, вы узнаете, как автоматизировать извлечение синонимов из данных поисковой системы, избегая ручного редактирования файлов синонимов (глава 2). Это экономит время, повышая эффективность поиска, независимо от конкретного случая использования или области. То же самое относится и к подходящим предложениям по сопутствующему контенту (глава 6). Во многих случаях пользователи удовлетворены сочетанием простого поиска с возможностью навигации по сопутствующему контенту. Мы также рассмотрим некоторые более конкретные случаи использования, такие как поиск контента на нескольких языках (глава 7) и поиск изображений (глава 8).

Единственное требование к методам, которые мы будем обсуждать, заключается в том, что у них достаточно данных для подачи в нейронные сети. Но в общем виде трудно определить границы «достаточного количества данных». Вместо этого давайте суммируем минимальное количество документов (текст, изображения и т. д.), которые обычно необходимы для каждой проблемы, рассматриваемой в книге: *см*. табл. 1.1.

Таблица 1.1. Требования к задачам для методов поиска на базе нейронных сетей

Задача	**Минимальное количество документов (диапазон)**	**Глава**
Изучение представлений слов	1000–10 000	2, 5
Генерирование текста	10 000–100 000	3, 4
Изучение представлений документов	1000–10 000	6
Машинный перевод	10 000–100 000	7
Изучение представлений изображений	10 000–100 000	8

Обратите внимание, что не нужно строго следовать этой таблице; цифры взяты из опыта. Например, даже если в поисковой системе насчитывается менее 10 000 документов, вы все равно можете попытаться реализовать методы нейронного машинного перевода, описанные в главе 7; но вы должны принять во внимание тот факт, что получить качественные результаты может быть труднее (например, совершенные переводы).

Читая книгу, вы многое узнаете о глубоком обучении, а также обо всех необходимых основах поиска для реализации этих принципов ГО в поисковой системе. Поэтому, если вы инженер, связанный с разработкой поисковых систем, или программист, который хочет изучать поиск на базе нейронных сетей, эта книга для вас.

На данный момент вам не обязательно знать, что такое глубокое обучение или как оно работает. Вы подробнее узнаете об этом, когда мы будем рассматривать