ГЛАВА 1

Обзор GPT-4 и ChatGPT

Сегодня разработчикам предоставляется шанс раскрыть потенциал искусственного интеллекта. *Большие языковые модели* (large language models, LLM), такие как GPT-4 и GPT-3.5 Turbo, уже продемонстрировали свои возможности в ChatGPT. Сейчас мы находимся в вихре прогресса, темпы которого никогда ранее не наблюдались в мире программного обеспечения (ПО). Компания OpenAI сделала эти технологические инновации легкодоступными. Какие преобразующие приложения вы создадите с помощью имеющихся в вашем распоряжении инструментов?

Способности этих ИИ-моделей выходят далеко за рамки чат-ботов. Благодаря LLM разработчики теперь могут использовать преимущества *обработки естественного языка* (natural language processing, NLP) для создания приложений, которые понимают потребности пользователей, превращая то, что когда-то было научной фантастикой, в осязаемую реальность. Более того, благодаря новым алгоритмам работы с изображениями GPT-4 теперь можно создавать ПО, которое способно интерпретировать и генерировать текст на основе изображений и снимков. Языковые модели GPT открывают совершенно новый мир возможностей: от инновационных систем клиентской поддержки, которые учатся и адаптируются, до персонализированных образовательных инструментов, понимающих уникальный стиль обучения каждого студента.

Но что *представляют собой* эти модели GPT? В первой главе мы изучим их основы, происхождение и ключевые особенности. Поняв суть этих моделей искусственного интеллекта, вы уже будете на пути к созданию следующего поколения приложений, основанных на LLM.

Введение в LLM

В этом разделе мы заложим фундамент, на котором строятся модели GPT. Наша цель — дать вам полное представление о языковых моделях и NLP, роли архитектуры трансформеров, а также о процессах токенизации и прогнозирования в GPT-моделях. Однако, как вы увидите, наше путешествие не ограничивается обработкой текста. Появление GPT-4 Vision расширяет возможности LLM за пределы текста и включает в себя обработку *мультимодального* ввода. Это означает, что GPT-4 умеет не только обрабатывать текст, но и интерпретировать изображения.

Изучение основ языковых моделей и NLP

Как и LLM, GPT представляют собой новейший тип моделей, появившихся в области NLP — одном из направлений в сфере *машинного обучения* (machine learning, ML) и искусственного интеллекта. Прежде чем перейти к рассмотрению моделей GPT, необходимо взглянуть на NLP и смежные с ним области.

Существуют различные определения ИИ, но одно из самых популярных гласит, что искусственный интеллект — это компьютерные системы, которые могут выполнять задачи, обычно требующие человеческого интеллекта. Согласно такому определению, многие алгоритмы можно отнести к ИИ: например, алгоритм прогнозирования трафика в GPS-приложениях или системы на основе правил, используемые в стратегических видеоиграх. Со стороны может показаться, что для выполнения этих задач машине требуется интеллект.

Машинное обучение — это подмножество искусственного интеллекта. В рамках ML мы не стремимся напрямую реализовать правила принятия решений, используемые системой ИИ. Наша цель — разработать алгоритмы, которые позволят системе самостоятельно обучаться на примерах. С 1950-х годов, когда начались исследования в сфере ML, в научной литературе было предложено множество алгоритмов ML.

Среди них особо выделяются алгоритмы глубокого обучения. *Глубокое обучение* (deep learning, DL) — это направление в области машинного обучения, которое фокусируется на алгоритмах, работающих наподобие человеческого мозга. Такие алгоритмы называются *искусствен-*

ными нейронными сетями. Они способны эффективно обрабатывать очень большие объемы данных и успешно решать задачи вроде распознавания изображений и речи, а также NLP.

Модели GPT основаны на архитектуре трансформеров *(Transformer)*, представленной в статье 2017 года *Attention Is All You Need* (https://oreil.ly/jVZW1) за авторством Васвани и др. из Google. Трансформеры похожи на читающие машины. Они используют механизм внимания для приоритизации различных частей текста, что позволяет глубже понять контекст и получить связные результаты. Такой подход позволяет им улавливать смысл слов в предложениях, что повышает их производительность при переводе, ответах на вопросы и создании текстов. На рис. 1.1 наглядно представлены основные концепции и их роль в расширении возможностей моделей-трансформеров для решения различных языковых задач.

Рис. 1.1. Вложенный набор технологий от искусственного интеллекта до трансформеров

NLP — это область искусственного интеллекта, позволяющая компьютерам обрабатывать, интерпретировать и генерировать естественный

человеческий язык. Современные решения в сфере NLP основаны на алгоритмах машинного обучения. NLP охватывает широкий круг задач.

- *Классификация текстов* по заранее определенным группам. К ним относятся, например, анализ настроений и категоризация тем. Компании могут использовать анализ настроений, чтобы определить мнение клиентов о своих услугах. Фильтрация электронной почты также является примером категоризации по темам, при которой письма могут быть отнесены к таким категориям, как «Личные», «Социальные», «Рекламные акции» и «Спам».
- *Автоматический перевод* текста с одного языка на другой. Сюда можно отнести не только перевод обычных текстов, но и перевод кода с одного языка программирования на другой, например с Python на C++.
- *Ответы на вопросы* на основе заданного текста. Например, онлайн-портал обслуживания клиентов может использовать модель NLP для ответа на часто задаваемые вопросы о продукте, а образовательное ПО — для предоставления ответов на вопросы студентов по изучаемой теме.
- *Генерация текста*. Формирование связного и релевантного выходного текста на основе заданного входного текста, называемого запросом или промтом (prompt).

Как вы уже знаете, LLM — это модели машинного обучения, предназначенные для решения, в частности, задач генерации текстов. Языковые модели позволяют компьютерам обрабатывать, интерпретировать и генерировать человеческий язык, обеспечивая более эффективное взаимодействие между человеком и машиной. Для достижения этой цели LLM анализируют огромные объемы текстовых данных или *обучаются* на них, выявляя закономерности и изучая связи между словами в предложениях. Для такого обучения могут использоваться различные источники данных: тексты из Википедии, Reddit, архивов книг или даже из самого Интернета. При наличии входного текста LLM учится предугадывать наиболее вероятные последующие слова и таким образом генерировать осмысленные ответы на входной текст. Современные языковые модели, появившиеся в последние несколько месяцев, настолько мощные и обучены на таком большом количестве текстов, что теперь они могут выполнять огромное количество задач NLP, включая классификацию текстов, машинный перевод, ответы на вопросы и многое другое. Модели GPT-4 и ChatGPT — это современные LLM, которые отлично справляются с задачами генерации текстов.

Компания OpenAI разработала различные языковые модели. На момент написания книги последними и наиболее мощными являются модели серии GPT-4. GPT-4 Vision представляет собой грандиозную мультимодальную модель, позволяющую работать как с текстом, так и с изображениями. LLM способны интерпретировать изображения, используя специализированную архитектуру трансформера, называемую *визуальным трансформером* (vision transformer, ViT). Совсем недавно модель GPT-4o пошла дальше в мультимодальности: она может обрабатывать и генерировать текст, изображение и звук.

В 1990-х годах стали развиваться LLM. Все началось с простых языковых моделей, таких как *n-граммы*, которые пытались «предсказать» следующее слово в предложении на основе предыдущих. Для этого в *n*-граммных моделях используется *частота*. Прогнозируемое слово — это наиболее часто встречающееся слово, следующее за предыдущими в тексте, на котором обучалась модель. Хотя подобный подход казался неплохим началом, качество понимания контекста и грамматики у *n*-граммных моделей оставляло желать лучшего, что привело к проблемам с генерацией согласованного текста.

Для повышения производительности *n*-граммных моделей были разработаны более сложные алгоритмы обучения, в том числе рекуррентные нейронные сети (recurrent neural networks, RNN) и сети с долгой краткосрочной памятью (long short-term memory, LSTM). Эти модели могли обучаться более длинным последовательностям и анализировать контекст лучше, чем *n*-граммы, однако им по-прежнему требовалась помощь для эффективной обработки больших объемов данных. Долгое время именно эти типы рекуррентных моделей считались наиболее эффективными и поэтому часто применялись в таких задачах, как автоматический машинный перевод.

Ключевые особенности трансформера и его роль в обработке естественного языка

Трансформеры совершили революцию в NLP, прежде всего благодаря тому, что успешно преодолели одно из критических ограничений предыдущих моделей NLP — рекуррентных нейронных сетей, а именно их неспособность работать с длинными текстовыми последовательностями и запоминать контекст при таких объемах. Другими словами, если

RNN склонны забывать контекст при больших объемах информации (печально известное «катастрофическое забывание»), то трансформеры способны эффективно обрабатывать и сохранять его.

Ключевым элементом этой технологии является *механизм внимания* — простая, но мощная идея. Вместо того чтобы рассматривать все слова в последовательности текста как одинаково важные, модель «обращает внимание» на наиболее релевантные термины для каждого шага своей задачи. Этот механизм позволяет устанавливать прямые связи между удаленными друг от друга элементами текста, так что последнее слово может «присутствовать» в первом слове без каких-либо условий, преодолевая существенное ограничение, с которым сталкивались предыдущие модели, такие как RNN. Перекрестное (или кросс-) внимание и самовнимание — два метода, которые основаны на механизме внимания и часто встречаются в LLM, в частности в архитектуре трансформера.

Перекрестное внимание (cross-attention) помогает модели определить значимость различных частей входного текста для точного предсказания следующего слова в выходном тексте. Представьте прожектор, который светит на слова или фразы во входном тексте, выделяя релевантную информацию, необходимую для предсказания следующего слова, и игнорируя менее важные детали.

Для наглядности рассмотрим пример с переводом простого предложения. Допустим, у нас есть предложение на английском языке *Alice enjoyed the sunny weather in Brussels*, которое необходимо перевести на французский: *Alice a profité du temps ensoleillé à Bruxelles*. В примере мы сосредоточимся на генерации французского слова *ensoleillé*, которое означает *sunny* («солнечный»). При прогнозировании механизм перекрестного внимания присвоит больший вес английским словам *sunny* и *weather*, поскольку они оба имеют отношение к значению слова *ensoleillé*. Фокусируясь на этих двух словах, перекрестное внимание помогает модели создать точный перевод части предложения (рис. 1.2).

Под *самовниманием* (self-attention) понимается способность модели фокусироваться на различных частях входного текста. В контексте NLP это позволяет модели оценивать значимость каждого слова в предложении по отношению к другим. Это помогает ей лучше разбираться в связях между словами и формировать новые *концепции* на основе нескольких слов во входном тексте.

В качестве более конкретного примера можно привести следующее предложение: «Алиса получила похвалу от своих коллег». Предполо-

жим, что модель пытается понять значение слова *«своих»*. Механизм самовнимания присваивает различные веса словам в предложении и выделяет те, которые относятся к слову «своих» в данном контексте. В нашем случае больший вес придается словам *«Алиса»* и *«коллеги»*. Самовнимание помогает модели строить новые понятия на основе этих слов. В данном примере одним из понятий, которое может появиться, будет «коллеги Алисы», как показано на рис. 1.3.

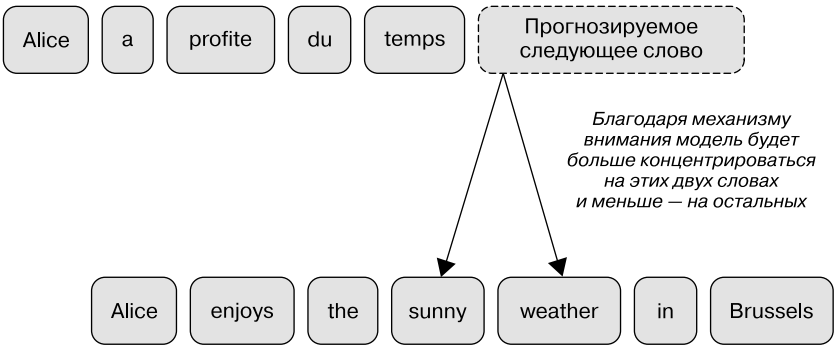

Рис. 1.2. Перекрестное внимание использует механизм внимания для фокусировки на существенных частях входного текста (английское предложение), чтобы предсказать следующее слово в выходном тексте (французское предложение)

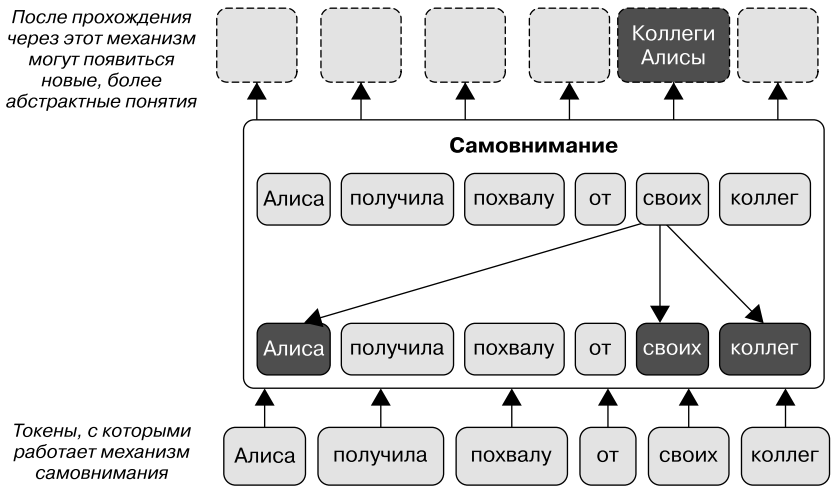

Рис. 1.3. Самовнимание стало причиной появления фразы «коллеги Алисы»

В отличие от рекуррентной архитектуры, трансформеры обладают дополнительным преимуществом — их легко *распараллелить*. То есть трансформеры способны обрабатывать несколько частей входного текста одновременно, а не последовательно. Это позволяет ускорить вычисления и процесс обучения, поскольку различные части модели могут работать параллельно, не дожидаясь завершения предыдущих шагов, в отличие от рекуррентных архитектур, где требуется последовательная обработка. Возможность параллельной обработки трансформерных моделей идеально вписывается в архитектуру графических процессоров (GPU), предназначенных для одновременной обработки нескольких вычислений. Благодаря своей высокой параллельности и вычислительной мощности графические процессоры отлично подходят для обучения и запуска трансформерных моделей. Это достижение позволило специалистам по анализу данных обучать модели на гораздо больших наборах данных, что стало ключевым моментом на пути к развитию LLM.

Архитектура трансформера — это модель последовательности, которая изначально разрабатывалась для решения задач преобразования последовательности в последовательность, таких как машинный перевод. Стандартный трансформер состоит из двух основных компонентов: кодировщика и декодировщика, оба из которых сильно зависят от механизма внимания. Задачами первого являются обработка входного текста, выявление ценных признаков и формирование осмысленного представления этого текста, которое называется *вложением,* или *эмбеддингом (embedding)*. Затем декодировщик использует это вложение для создания выходного результата, например перевода или резюме. В итоге получается эффективная интерпретация закодированной информации.

Генеративные, предварительно обученные трансформеры, обычно называемые GPT (generative pre-trained transformers), представляют собой семейство моделей, основанных на архитектуре трансформеров и специально использующих часть оригинальной архитектуры, отвечающую за декодирование. В GPT кодировщик отсутствует, поэтому нет необходимости в механизме перекрестного внимания для интеграции вложений, произведенных кодировщиком. В результате GPT полагается исключительно на механизм самовнимания внутри декодировщика для создания контекстно зависимых представлений и прогнозов. Отметим, что другие известные модели, такие как BERT (Bidirectional Encoder Representations from Transformers), основаны на кодирующей части. Данный тип модели в книге не рассматривается. На рис. 1.4 показана эволюция различных моделей.

Рис. 1.4. Эволюция методов NLP от n-грамм до LLM

Суть процесса токенизации и прогнозирования в моделях GPT

Языковые модели в семействе GPT принимают на вход промт (запрос), а в ответ генерируют текст. Этот процесс известен как *завершение текста*. Например, запрос может звучать так: *«Сегодня хорошая погода, и я решил»*, а выходной текст модели — так: *«Пойти прогуляться»*. Вам может быть интересно, как модель LLM строит свой ответ на основе вашего ввода. Как вы увидите, это в основном просто вопрос вероятностей.

Когда запрос отправляется в LLM, вводимый текст сначала разбивается на более мелкие фрагменты, называемые *токенами*, которые представляют собой отдельные слова, части слов, пробелы и знаки пунктуации. Например, предыдущий запрос может быть разбит следующим образом: *[«Сегодня», «хорош», «ая», «пого», «да», «,», «по», «этому», «я», «ре», «шил»]*. У каждой языковой модели есть свой токенизатор. Токенизатор из серий GPT-3.5 и GPT-4 доступен для тестирования на платформе OpenAI (https://oreil.ly/hbKT7).

Следует отметить, что 100 токенов примерно равны 75 словам на английском языке. Однако для других языков это значение может быть другим.

Используя принцип внимания и архитектуру трансформеров, LLM обрабатывает эти токены и интерпретирует связи между ними и общий смысл запроса. Трансформеры позволяют модели эффективно идентифицировать в тексте критическую информацию и его контекст.

Для создания нового предложения LLM прогнозирует наиболее вероятные токены, исходя из контекста запроса. Компания OpenAI создала

две версии GPT-4 с контекстными окнами с количеством токенов 8192 и 32 768. На начало 2024 года последними моделями, выпущенными OpenAI, являются GPT-4 Turbo и GPT-4o, с более крупным входным контекстным окном в 128 000 токенов, что эквивалентно почти 300 страницам текста на английском языке. В отличие от предыдущих рекуррентных моделей, которые с трудом справлялись с обработкой длинных входных последовательностей, архитектура трансформеров с механизмом внимания позволяет современной LLM рассматривать контекст как единое целое. Основываясь на этом, модель присваивает вероятностные оценки каждому потенциальному последующему токену. Токен с наивысшей вероятностью выбирается в качестве очередного. В нашем примере после слов «Сегодня хорошая погода, и я решил» следующим по вероятности токеном может быть «пойти».

Как мы увидим в следующей главе, с помощью параметра `temperature` вместо одного токена с наибольшей вероятностью модель также может выбрать следующий токен из *набора* токенов с наибольшей вероятностью. Это позволяет добиться вариативности и креативности в ответе модели.

Затем этот процесс повторяется, но теперь контекст становится таким: «Сегодня хорошая погода, и я решил пойти», где к исходному предложению добавляется ранее предсказанный токен «пойти». Этот процесс повторяется до тех пор, пока не будет сформировано полное предложение: «пойти прогуляться». Данный метод основан на способности LLM учиться предугадывать следующее наиболее вероятное слово на основе массивных текстовых данных (рис. 1.5).

Интеграция видения в программу LLM

GPT-4 Vision расширяет возможности серии GPT-4, добавляя в нее мультимодальность и делая модель эффективной не только при работе с текстом. Конкретных механизмов, обеспечивающих эту функцию, компания OpenAI не раскрывает. Однако, чтобы получить представление о них и примерно понять, какие методики применяются в GPT-4 для достижения такой мультимодальной функциональности, достаточно изучить открытые LLM, которые интегрируют визуальные данные. В этом разделе мы рассмотрим процессы, наблюдаемые в открытых аналогах, чтобы разобраться, как интеграция изображения и текста может быть реализована в GPT-4.

Рис. 1.5. Процесс завершения текста является итерационным и выполняется токен за токеном

Сверточные нейронные сети (CNN) уже давно являются одной из передовых технологий для решения задач обработки изображений. CNN отлично справляются с классификацией изображений и обнаружением объектов за счет использования слоев фильтров, которые «скользят» по входному изображению. Эти фильтры могут поддерживать пространственные отношения между пикселями изображения. Благодаря этому слою фильтров CNN способны распознавать детали, начиная от простых краев в ранних слоях и заканчивая сложными формами и объектами в более глубоких.

Но подобно тому, как появление в 2017 году архитектуры трансформера произвело революцию в NLP, вытеснив RNN, в 2020 году были предложены новые модели на основе трансформеров для обработки изображений. С тех пор многолетнее господство CNN в области обработки изображений было поставлено под сомнение. В 2021 году в статье Google под названием *An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale* (https://oreil.ly/ijPSk) Досовицкий и др. показали, что чистая модель трансформера, названная визуальным трансформером (vision transformer, ViT), превосходит CNN в классификации изображений.