DEEPSEEK

0101010101010
1010101010100
1010101010100
1010101001010
1010100101010

ТЕХНОЛОГИИ ГЕНЕРАЦИИ

0101010101010
1010101010100
1010101010100
1010101001010
1010100101010

НЕЙРОСЕТЕВЫЕ

deepseek

ТЕКСТА

0101010101010
1010101010100

ЛЭЙЭНСТАЗИЯ

Лэй Энстазия

DeepSeek – нейросетевые технологии генерации текста

http://www.litres.ru/pages/biblio_book/?art=71578282
SelfPub; 2025

Аннотация

Книга посвящена инновационной языковой модели, которая меняет представление об искусственном интеллекте. DeepSeek сочетает обучение с подкреплением (RL) и технологию Retrieval-Augmented Generation (RAG), обеспечивая точность, адаптивность и доступ к актуальной информации в реальном времени. В книге рассматриваются уникальные особенности модели, такие как отказ от традиционного обучения с учителем, интеграция многоголовой скрытой внимательности (MLA) и дистилляция знаний для компактных устройств. Авторы подробно описывают её применение в науке, образовании, программировании, медицине и других сферах, акцентируя внимание на её экологичности, надежности и способности минимизировать ошибки. Особое внимание уделено перспективам развития модели, включая борьбу с галлюцинациями, повышение энергоэффективности и этические вызовы автоматизации. Книга станет незаменимым пособием для

разработчиков, исследователей и всех, кто интересуется будущим искусственного интеллекта.

Содержание

Глава 1. Введение 5
Глава 2. Технологическая основа DeepSeek 10
Глава 3. Дистилляция знаний в DeepSeek 15
Глава 4. Преимущества DeepSeek 18
Конец ознакомительного фрагмента. 20

Лэй Энстазия DeepSeek – нейросетевые технологии генерации текста

Глава 1. Введение

1.1. Краткий обзор языковых моделей и их значения

Языковые модели стали фундаментом современного искусственного интеллекта (ИИ). Они используются в задачах обработки естественного языка (NLP), таких как машинный перевод, генерация текстов, анализ настроений и автоматическое резюмирование. С момента появления первых моделей на основе нейронных сетей, таких как Word2Vec и GloVe, прогресс в этой области был стремительным.

Трансформеры (Transformer) стали ключевым прорывом, положив начало новой эре языковых моделей. Модели, такие как BERT, GPT и T5, научились решать сложные зада-

чи, учитывая контекст, долгосрочные зависимости и многослойные представления данных.

Появление таких моделей открыло новые возможности:

Улучшение автоматизации в бизнесе, науке и образовании.

Сокращение времени на обработку больших объемов информации.

Расширение доступа к технологиям, делающим сложные задачи доступными даже для небольших компаний.

Однако традиционные подходы, основанные на обучении с учителем (Supervised Fine-Tuning, SFT), показали свои ограничения.

1.2. Эволюция от SFT к RL и появление DeepSeek

Традиционное обучение с учителем предполагает использование заранее подготовленных данных, где модель учится ассоциировать входные данные с соответствующими ответами. Но этот подход имеет недостатки:

Ограниченная гибкость: Модели узко специализированы и плохо адаптируются к новым задачам.

Зависимость от качества данных: Ошибки или предвзятость в данных приводят к снижению точности модели.

Неспособность к обобщению: Модели, обученные на SFT, часто не способны справляться с невиданными ранее ситуациями.

DeepSeek стал революционным шагом благодаря использованию обучения с подкреплением (Reinforcement Learning, RL). В отличие от SFT, RL позволяет модели учиться, взаимодействуя с окружающей средой.

Основные преимущества RL:

Самообучение: Модель учится на своих ошибках, улучшая стратегию поведения.

Гибкость: RL помогает адаптироваться к изменениям и новым условиям.

Развитие цепочек рассуждений (Chain of Thought, CoT): Модель способна разбивать сложные задачи на этапы и находить оптимальные решения.

DeepSeek интегрировал RL на новом уровне, отказавшись от начального этапа SFT. Это позволило создать модель, которая не только решает задачи, но и эффективно адаптируется к новым вызовам.

1.3. Важность технологий RAG (Retrieval-Augmented Generation) для будущего моделей

Retrieval-Augmented Generation (RAG) – это подход, при котором языковая модель получает доступ к внешним источникам данных. Вместо того чтобы полагаться только на запомненные знания, модель активно запрашивает и использует актуальную информацию в реальном времени.

Технология RAG играет ключевую роль в будущей эволюции языковых моделей:

Повышение точности: Вместо запоминания устаревшей информации модель использует актуальные данные.

Масштабируемость: Объем знаний не ограничен параметрами модели.

Снижение риска галлюцинаций: Модель опирается на внешние источники, уменьшая вероятность генерации оши-

бочной информации.

DeepSeek успешно использует RAG, чтобы предложить пользователям наиболее точные и релевантные ответы, применяя RL для оптимизации запросов и выбора информации.

Заключение главы

Эта глава подводит читателя к пониманию, почему DeepSeek является не просто новым участником на рынке языковых моделей, но и значительным шагом вперед. Используя преимущества RL и RAG, DeepSeek задает новый стандарт для моделей следующего поколения.

Глава 2. Технологическая основа DeepSeek

2.1. Архитектура модели

DeepSeek использует усовершенствованную архитектуру на основе многоголовой скрытой внимательности (MLA), что позволяет значительно повысить производительность модели. Основные аспекты MLA в DeepSeek:

Разделение внимания: Каждая "голова" фокусируется на разных аспектах данных – семантике, синтаксисе или контексте. Это улучшает способность модели улавливать глубокие связи внутри данных.

Параллельная обработка: MLA реализует одновременное вычисление для каждой головы, что сокращает время обработки и делает модель масштабируемой.

Инновации MLA: DeepSeek интегрирует оптимизацию вычислений за счет игнорирования несущественных данных (аналог сжатия контекста). Эта технология позволяет снизить затраты на обучение и уменьшить энергопотребление,

оставаясь при этом конкурентоспособной.

В сравнении с классическими моделями, такими как GPT и BERT, DeepSeek демонстрирует:

Более низкие затраты на вычисления, благодаря улучшенной оптимизации MLA.

Способность к обучению на меньшем объеме данных без потери точности.

Лучшую адаптацию к задачам, требующим актуальной и контекстной информации.

2.2. Отказ от SFT

DeepSeek полностью пересмотрел традиционный подход обучения с учителем (Supervised Fine-Tuning, SFT). Это решение обусловлено недостатками SFT:

Жесткая привязка к исходным данным: Модели SFT зависимы от качества подготовленного набора данных. Ошибки или предвзятость могут существенно снижать качество обучения.

Ограниченные возможности обобщения: Модели

SFT зачастую неспособны справляться с новыми задачами, которые не отражены в обучающей выборке.

Вместо этого DeepSeek делает ставку на обучение с подкреплением (RL):

Гибкость и адаптивность: Модель активно взаимодействует с окружающей средой, развивая навыки, основанные на результатах своих действий.

Развитие цепочек рассуждений (CoT): DeepSeek разбивает сложные задачи на этапы, оптимизируя каждый шаг для достижения лучшего результата.

Самообучение: Модель способна учиться на своих ошибках и корректировать поведение, что делает ее менее зависимой от начальных данных.

Этот переход позволил DeepSeek стать не просто инструментом, выполняющим команды, но и динамической системой, которая адаптируется к изменениям и новым вызовам.

2.3. Механизм RAG

Технология Retrieval-Augmented Generation (RAG) является ключевым элементом DeepSeek. В отличие от традици-

онных моделей, которые полагаются только на запомненные знания, RAG позволяет интегрировать актуальную информацию из внешних источников в процессе работы модели.

Основные преимущества RAG в DeepSeek:

Реальный доступ к данным: Модель может находить и использовать релевантные сведения из актуальных баз знаний, что снижает риск генерации устаревших или неверных ответов.

Улучшение контекстуальности: DeepSeek эффективно связывает вводимые данные пользователя с внешними источниками, адаптируя свои ответы под конкретную задачу.

Оптимизация запросов: С помощью RL DeepSeek обучается формировать максимально точные запросы к базам знаний, что ускоряет получение релевантных данных.

Примеры реализации RAG в DeepSeek:

Генерация научных текстов с использованием последних публикаций.

Автоматизация юридических анализов, включая поиск законов и судебных решений.

Обеспечение поддержки пользователей с доступом к обновленным техническим инструкциям.

Заключение главы

DeepSeek объединяет лучшие технологии в области языковых моделей – MLA для оптимизации обработки данных, RL для адаптивного обучения и RAG для интеграции актуальной информации. Эта комбинация делает модель универсальным инструментом, способным решать широкий спектр задач в реальном времени.

Глава 3. Дистилляция знаний в DeepSeek

3.1. Как перенос знаний улучшает производительность

Дистилляция знаний (Knowledge Distillation) – это процесс передачи опыта от большой, высокопроизводительной модели ("учителя") к более компактной модели ("ученику"). DeepSeek использует этот подход для достижения баланса между производительностью и экономичностью.

Основные аспекты:

Концентрация ключевых знаний: Вместо передачи всех данных модель "учитель" фокусируется на важных аспектах, таких как вероятности или представления значимых взаимосвязей.

Ускорение обучения: Модель "ученик" быстрее достигает желаемой производительности благодаря эффективному обучению на основе выверенных выходных данных "учителя".

Сохранение высокого качества: Дистиллированные модели демонстрируют сопоставимые результаты с оригинальными большими моделями, сохраняя точность и контекстуальность.

3.2. Компактные модели для устройств с ограниченными ресурсами

DeepSeek решает важную задачу – расширение возможностей компактных моделей для использования на устройствах с ограниченной вычислительной мощностью.

Преимущества компактных моделей:

Доступность: Они легко интегрируются в мобильные устройства, IoT-системы и локальные серверы.

Снижение энергопотребления: Оптимизированные модели требуют меньше вычислительных ресурсов, что делает их подходящими для экологичных и масштабируемых решений.

Расширение областей применения: Компактные версии DeepSeek позволяют внедрять ИИ даже в тех местах, где ранее это было невозможно, например, в сельских регионах

или автономных системах.

Пример: локальная модель DeepSeek может использоваться врачами в удаленных районах для мгновенного анализа медицинских данных без необходимости подключения к облачным системам.

Глава 4. Преимущества DeepSeek

4.1. Высокая производительность и демократизация доступа к ИИ

DeepSeek сочетает передовые алгоритмы с продуманной оптимизацией, что позволяет ей достигать результатов, сравнимых с лидерами отрасли, такими как GPT-4, при более низких затратах.

Доступность для разработчиков: Открытый исходный код DeepSeek и поддержка сообществом способствуют созданию новых решений на ее основе.

Снижение барьеров входа: Низкая стоимость использования делает ИИ доступным для малого и среднего бизнеса, а также образовательных и исследовательских учреждений.

Поддержка инклюзивности: DeepSeek нацелен на распространение ИИ в глобальном масштабе, независимо от экономического состояния региона.

4.2. Снижение стоимости обучения и использования

Инновации в архитектуре DeepSeek и ее обучение с использованием RL приводят к снижению вычислительных затрат.

Ключевые аспекты:

Оптимизация обучения: Использование "холодного старта" и RL вместо массивных наборов данных для SFT.

Энергоэффективность: Значительное снижение потребления электроэнергии, что делает DeepSeek более экологичным выбором.

Масштабируемость: Гибкость DeepSeek позволяет развертывать модель с минимальными затратами даже на инфраструктуре с ограниченными ресурсами.

Конец ознакомительного фрагмента.

Текст предоставлен ООО «Литрес».

Прочитайте эту книгу целиком, купив полную легальную версию на Литрес.

Безопасно оплатить книгу можно банковской картой Visa, MasterCard, Maestro, со счета мобильного телефона, с платежного терминала, в салоне МТС или Связной, через PayPal, WebMoney, Яндекс.Деньги, QIWI Кошелек, бонусными картами или другим удобным Вам способом.