Как теория и практика встречаются в современных нейронных сетях

Соколов Евгений Андреевич

03.07.2025

Изначально задача машинного перевода решалась по аналогии с человеческим подходом: через изучение слов и грамматики. Это привело к созданию систем, основанных на правилах, но такой метод оказался ограниченным — невозможно было вручную закодировать все нюансы языка. Тогда появилась идея отказаться от жестких правил и вместо этого использовать данные: эксперты предоставляли примеры переводов, а алгоритм подбирал параметры сложной модели, чтобы максимизировать качество. Поскольку параметров было очень много, потребовались не только значительные вычислительные ресурсы, но и эффективные алгоритмы оптимизации. В результате нейросетевые модели могут научиться делать качественный перевод всего за несколько недель обучения.

Специализированные модели создаются под конкретные задачи, такие как машинный перевод, поиск, рекомендательные системы или распознавание речи. Их ключевые особенности заключаются в том, что они обучаются на узкоспециализированных данных, решают строго определённую проблему и могут выдавать огромное количество результатов в своей области. Однако их главный недостаток — узкая специализация: они не способны решать задачи за пределами своей сферы.

В отличие от специализированных, *Foundation Models* решают общую задачу — например, предсказание следующего слова в тексте. Они используют схожие методы оптимизации, но требуют гораздо больше данных и вычислительных мощностей. Такие модели не ограничены одной областью и могут адаптироваться к разным задачам.

Несмотря на развитие универсальных моделей, специализированные решения остаются востребованными. Например, нейросети, обученные генерировать код, помогают людям без навыков программирования создавать программы. Интересно, что нейросети способны решать даже геометрические задачи, которые традиционно требуют визуального мышления. Для формальной записи и проверки таких решений был разработан специальный язык, на котором можно автоматически проверять решение, что открывает новые возможности автоматизации в математике.

Современные языковые модели обучаются в два этапа. На первом этапе модель обучается на обширных данных из интернета, где получает все основные знания. Второй этап представляет собой тонкую настройку, где модель дообучается на тщательно отобранных данных, что значительно улучшает корректность её ответов.

Закончились ли данные для обучения? Хотя крупнейшие модели уже обучены на большей части текстового интернета, есть пути дальнейшего развития, например, планируется добавление новых модальностей, т.е. обучение на изображениях, видео, аудиоматериалах и коде).

Основу современных LLM составляет архитектура *трансформеров (Transformer)*, которая использует механизм внимания для обработки длинных последовательностей. Чем больше параметров у модели (сотни миллиардов и более), тем лучше она справляется с задачами.

Scaling Laws (законы масштабирования) — эмпирически обнаруженная зависимость, заключающаяся в том, что качество модели линейно улучшается с ростом объема данных, вычислительных ресурсов и числа параметров, однако после определенного предела рост замедляется, требуя новых оптимизаций.

Следующий шаг — создание *рассуждающих моделей (Reasoning Models)*, которые не просто предсказывают слова, а выстраивают логические цепочки. Чем больше "размышлений" (цепочка промежуточных шагов), тем точнее ответ.

ИИ-агенты — модели, способные взаимодействовать с внешними сервисами (поиск, API, базы данных), например ReAct — подход, где модель чередует рассуждения ("мысль") и действия (запрос к инструменту).

Критические параметры (супервеса) В больших языковых моделях существуют особые 5-6 параметров из сотен миллиардов, которые играют ключевую роль. При их изменении vодель полностью ломается, нарушается распределение вероятностей слов, из-за чего увеличивается генерация артиклей и других часто используемых слов. Интересно, что увеличение этих параметров может улучшить качество генерации. Это особенно важно учитывать при квантизации (сжатии значений параметров), так как уменьшение супервесов ухудшит качество генерации.

Где трансформеры хранят данные? Хотя от языковых моделей ожидают общих знаний и преобразования текста, возникает вопрос - можно ли изменить конкретные данные внутри модели? Эксперименты показывают, что это возможно. Создается база фактологических вопросов с правильными ответами. Трансформер состоит из повторяющихся блоков обработки слов и внутри каждого блока работает механизм внимания (обмен информацией между словами), а

полносвязные слои добавляют своеобразный "шум" к обработке. Можно точно определить место в полносвязном слое, где формируется конкретный ответ и возможно настроить блок для выдачи заранее заданного неправильного ответа. Такие изменения трудно обнаружить, что создает проблемы с проверкой целостности модели.

Архитектура *Mixture of Experts* предлагает решение через распределение задач между специализированными модулями, что может повысить качество работы.

Таким образом, изучая нейросети, можно повысить их качество.