ChatGPT

1.2 Движение истории: почему сейчас?

Лимиты до-трансформерных моделей

До 2017 года прогресс в обработке языка опирался на **рекуррентные нейронные сети** (RNN) и их производную — **длинную краткосрочную память (LSTM)**. Эти модели обрабатывали последовательность по очереди — «слово за словом» — и могли терять контекст, который появлялся позже в предложении. Google исследователь Яков Ушкореит сравнивал LSTM с пластырем: метод позволял обрабатывать большие последовательности, но всё равно проходил текст последовательно, из-за чего «мы не могли извлечь достаточно информации для работы в масштабе» [1]. Кроме того, как отмечает обзор по трансформерам, RNN обрабатывают вход последовательно, что **не позволяет полноценно использовать GPU и делает обучение медленным**; они также плохо справляются с длинными зависимостями, потому что информация передается от шага к шагу и теряется на длинных цепочках [2]. Эти ограничения тормозили применение нейросетей на задачах машинного перевода, суммирования и диалога.

Изобретение механизма внимания и создание Transformer

Идея **самовнимания (self-attention)** возникла у Ушкореита около 2014 года. В отличие от RNN, этот механизм позволяет каждому слову сопоставлять себя с любой другой частью текста. Самовнимание «считывает всё и даёт более эффективный способ смотреть на много входов одновременно» [3]. Ушкореит считал, что такая модель будет **быстрее и эффективнее** рекуррентных сетей, потому что позволяет проводить вычисления параллельно на специализированных чипах [3].

В 2016 году он начал убеждать коллег в перспективности самовнимания, но многие считали, что идея **отказа от рекуррентных сетей** — еретическая [4]. Постепенно к проекту присоединились разработчики Иллия Полосухин, Ашиш Васвани и Ники Пармар. Они разработали архитектуру под названием **«Transformers: Iterative Self-Attention and Processing for Various Tasks»** — названием они хотели подчеркнуть, что модель *преобразует* входные данные, извлекая из них смысл [5]. Команда создала две версии: базовую, обученную за 12 часов, и модель **Big**, которая обучалась три с половиной дня [6]. Результаты превзошли конкурентов: базовая модель показала лучшую точность и при этом требовала меньше времени на обучение; модель Big побила существующие рекорды BLEU и её показатели продолжали улучшаться [7]. Сразу после этого они добавили в работу фразу, что планируют применять трансформер к другим модальностям, таким как **изображения, аудио и видео**, — намёк на будущую генеративную революцию [8].

Технически Transformer основывается **только на механизме внимания**, полностью отказываясь от рекуррентных и сверточных слоёв. Авторы подчеркнули, что такая архитектура **параллелизируется** лучше и требует значительно меньше времени на обучение, при этом достигая рекордных результатов на задачах машинного перевода [9].

Масштабы данных и вычислений: катализатор взрыва

Первые эксперименты показали, что самовнимание работает не хуже LSTM, но команда лишь нащупывала потенциал [10]. Ситуация изменилась, когда к проекту присоединился опытный инженер **Ноам Шазир**. Он посчитал существующие рекуррентные сети «раздражающими» и призвал «заменить их» [11]. Благодаря его оптимизациям и опыту, архитектура стала значительно проще и эффективнее. Команда провела серию **абляций**, постепенно убирая компоненты, чтобы понять, какие действительно нужны. Итогом стала **минималистичная** структура, обеспечивающая высокие результаты [7].

Важный фактор — **масштаб данных и вычислений**. Трансформер позволяет эффективно распараллеливать обучение на кластерах GPU и TPU, что открывает путь к обучению на многоязычных корпусах и соединению множества доменов. В статье Карпаты *«Software 2.0»* отмечается, что переход к нейронным методам делает обучение **компиляцией данных**, а разработчики теперь управляют датасетами [12]. Возможность собирать огромные наборы текстов и обучать модель на них, а затем масштабировать архитектуру до сотен миллиардов параметров, стала возможна только после появления трансформера. Именно поэтому сейчас, а не десять лет назад, мы наблюдаем взрыв генеративных моделей.

Распространение и эффект на индустрию

Публикация «Attention Is All You Need» в 2017 году не сразу перевернула мир. Даже внутри Google предлагалось заменить поисковый индекс трансформерами, но эту идею считали «сумасшедшей» [13]. Статья вышла летом, а в декабре на конференции NIPS команда получила шесть минут на презентацию. Однако стартап OpenAI быстро увидел потенциал: уже через несколько месяцев исследователи OpenAI использовали трансформер для создания первых прототипов GPT [13]. Google тоже сделал вклад, выпустив в 2018 году языковую модель **BERT**, основанную на трансформере [14]. Спустя несколько лет именно трансформер стал основой таких продуктов, как ChatGPT, DALL-E и другие генеративные системы [15].

Сегодня сложно представить область, где трансформер не используется. Его универсальность позволяет объединять разные модальности — текст, код, изображения — под одной архитектурой [16]. В результате мы наблюдаем так называемый **взрыв LLM**: комбинация архитектуры, масштабируемого обучения и гигантских корпусов данных создала инструменты, способные генерировать осмысленный текст, писать код и помогать в анализе данных.

Краткий сценарий доклада (≈ 5 мин)

1. **Предыстория: ограничения RNN** — объясните, почему рекуррентные сети не могли справляться с длинными текстами и зависели от последовательной обработки. Ушкореит называл LSTM «пластырем», который не даёт «правильного масштаба» [1].
2. **Рождение самовнимания** — расскажите, как Ушкореит и коллеги предложили механизмы самовнимания, дающие возможность каждому слову обращаться к любому другому слову. Эта идея позволяет распараллеливать вычисления и эффективнее использовать GPU [3].
3. **Создание Transformer** — опишите формирование команды и создание первых моделей. Подчеркните, что базовая версия уже опережала конкурентов, а модель Big побила рекорды BLEU и оказалась проще и быстрее других подходов [7]. Упомяните отказ от рекуррентных слоёв и сверточных операций [9].

4. **Масштаб данных и вычислений** — отметьте, что трансформер эффективен только при наличии больших данных и вычислительных мощностей. Возможность параллельного обучения и последующее увеличение параметров привели к появлению крупных языковых моделей. Это объясняет, почему взрыв генеративного ИИ случился именно сейчас.
5. **Распространение и влияние** — кратко расскажите о путешествии трансформера от внутреннего проекта Google до основного элемента ChatGPT, BERT и других систем. Укажите, что OpenAI и другие компании первыми реализовали потенциал, а гиганты вроде Google понадобились годы, чтобы довести технологии до продуктов [13].

Вывод

«Движение истории» в мире ИИ объясняется сочетанием нескольких факторов: появлением **трансформерной архитектуры**, уходом от рекуррентных ограничений, доступностью огромных корпусов данных и мощных кластеров GPU/TPU. Именно это сочетание позволило обучать модели, способные не просто классифицировать или распознавать, а **генерировать связные тексты, писать код и творчески решать задачи**. Понимание причин этого взрыва помогает осознанно использовать новые технологии и готовиться к дальнейшим революциям.

---

[1] [3] [4] [5] [6] [7] [8] [10] [11] [13] [14] [15] [16] 8 Google Employees Invented Modern AI. Here's the Inside Story | WIRED
https://www.wired.com/story/eight-google-employees-invented-modern-ai-transformers-paper/

[2] How Transformers Work: A Detailed Exploration of Transformer Architecture | DataCamp
https://www.datacamp.com/tutorial/how-transformers-work

[9] [1706.03762] Attention Is All You Need
https://arxiv.org/abs/1706.03762

[12] Software 2.0. I sometimes see people refer to neural... | by Andrej Karpathy | Medium
https://karpathy.medium.com/software-2-0-a64152b37c35