УДК 004.8

В. О. ГУЛЯЕВ, О. И. ЗАХАРОВА

ЭВОЛЮЦИЯ АРХИТЕКТУР ТРАНСФОРМЕРОВ: ОТ САМОВНИМАНИЯ К ЭФФЕКТИВНЫМ SPARSE-МОДЕЛЯМ

Поволжский государственный университет телекоммуникаций и информатики
г. Самара, Российская Федерация

Аннотация. *Статья посвящена исследованию новых архитектур искусственных нейронных сетей, направленных на улучшение классических трансформеров путем уменьшения вычислительных затрат и увеличения эффективности. Рассматриваются основные этапы эволюции трансформеров, начиная с введения механизма самовнимания и заканчивая современными sparse-моделями. Особое внимание уделяется методам оптимизации, таким как параметрически эффективный fine-tuning (PEFT), адаптивные слои (adapters), а также использование внешних хранилищ знаний и тет-векторов для расширения долговременной памяти моделей. Статья завершается обсуждением текущих достижений, ограничений и перспективных направлений будущих исследований в данной области.*

Ключевые слова: *трансформеры, механизм самовнимания, sparse-модели, PEFT, тет-векторы*

Введение

Трансформеры стали одной из самых влиятельных архитектур в области искусственного интеллекта, обеспечив значительное повышение точности и скорости обработки последовательных данных. Их успех обусловлен введением механизма самовнимания, который позволяет модели параллельно обрабатывать все элементы последовательности, выявляя важные связи между ними. Несмотря на свои достоинства, классические трансформеры сталкиваются с проблемой высоких вычислительных затрат, особенно при увеличении длины входных данных. Для решения этой проблемы исследователи разработали ряд новых подходов, включая sparse-модели и методы параметрически эффективного дообучения (PEFT).

Согласно данным Serverflow архитектура трансформеров основана на принципе самовнимания, что обеспечивает им преимущество перед предыдущими архитектурами, такими как рекуррентные нейронные сети (RNN). Однако рост числа параметров и вычислительная сложность делают традиционные трансформеры ресурсоемкими. Поэтому возникла необходимость разработки более эффективных архитектур.

Исторический обзор трансформеров

Классические трансформеры, представленные в статье «Attention Is All You Need» в 2017 году, положили начало революции в области обработки естественных языков (NLP). Трансформеры кардинально отличаются от предыдущих поколений нейросетей своей способностью одновременно рассматривать всю последовательность ввода. Благодаря механизму самовнимания каждая позиция входа получает уникальную оценку важности относительно остальных позиций, формируя динамические отношения между элементами данных. Этот подход устраняет проблему долгосрочной зависимости, характерную для RNN, и обеспечивает быстрое параллельное выполнение вычислений.

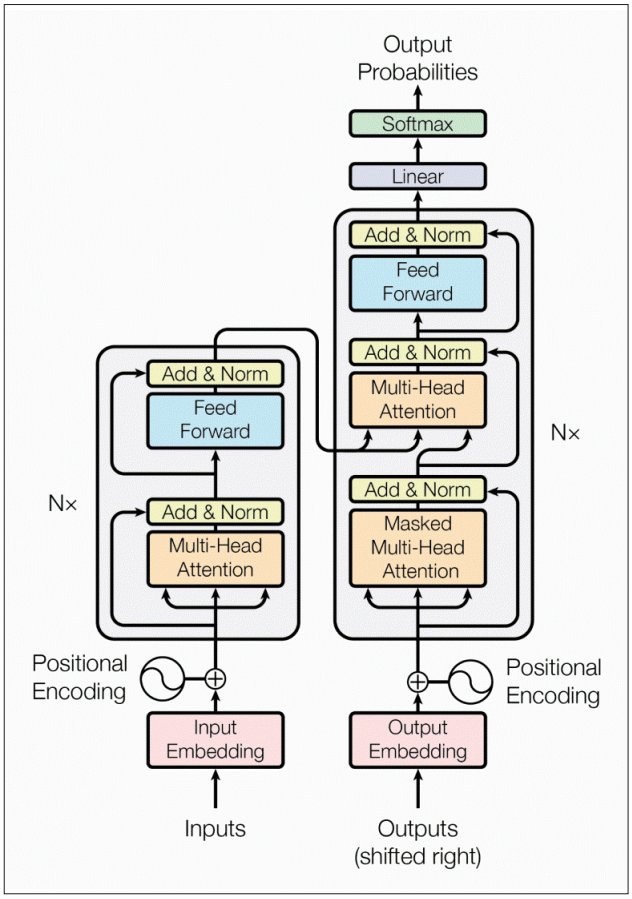

Рисунок 1. Архитектура Transformer

Механизм самовнимания основывается на трех ключевых компонентах: ключах («keys»), запросах («queries») и значениях («values»). Каждый элемент входящего текста преобразуется в три вектора, которые взаимодействуют друг с другом для формирования взвешенного представления контекста. Таким образом, трансформеры создают мощный инструмент для понимания семантической структуры данных.

Однако классическая архитектура имеет и недостатки. Основной недостаток заключается в экспоненциальном росте вычислительной нагрузки при увеличении размера вводимых данных. Время выполнения операций растет пропорционально квадрату длины последовательности, что создает трудности при применении моделей на длинных документах или временных рядах [1].

Современные подходы к повышению эффективности трансформеров

Учитывая перечисленные выше проблемы классических трансформеров, научное сообщество активно ищет пути повышения эффективности этих моделей. Следующим этапом рассмотрим современные подходы, направленные на снижение вычислительной сложности и улучшение масштабируемости трансформеров, что позволит расширить область их применения и повысить практическую ценность. Чтобы справиться с ограничениями классических трансформеров, исследователями были предложены различные инновационные подходы, направленные на повышение эффективности и снижение вычислительных затрат. Рассмотрим подробнее некоторые из наиболее успешных современных методов.

Sparse-модели

Классические трансформеры сталкиваются с серьезным ограничением: объем используемой памяти растет квадратично с увеличением длины последовательности. Причина – в механизме внимания, который вычисляет веса для всех пар токенов, формируя плотную матрицу. Sparse-модели трансформеров решают эту проблему иначе: они фокусируются только на наиболее значимых взаимодействиях, создавая разреженную матрицу внимания, где явно задано лишь «$n * n$» элементов.

Благодаря этому модели могут эффективно обрабатывать очень длинные контексты. Яркий пример – генерация изображений, где учитываются сложные корреляции между пикселями. Но есть и обратная сторона: модель не «видит» те связи, которые исключены из разреженной структуры. Впрочем, даже токены без прямой связи способны влиять друг на друга через более глубокие слои – как в сверточных сетях, где рецепторное поле растет с добавлением новых слоев.

Еще одно преимущество подхода – возможность явно заложить в модель предположения о структуре данных (inductive bias). Мы описываем входные данные через сущности (слова, пиксели, узлы графа), а затем, проектируя структуру разреженной матрицы, указываем, какие из них, по нашему мнению, должны взаимодействовать. Остается лишь подобрать подходящую архитектуру и проверить, насколько хорошо модель обучается в таких условиях [2].

Параметрически эффективный fine-tuning (PEFT)

Параметрически эффективный fine-tuning направлен на преодоление ключевого ограничения больших языковых моделей (LLMs) – избыточной ресурсоемкости, обусловленной необходимостью обновления огромного числа параметров.

В традиционных схемах подвергаются все параметры модели (в объеме миллионов или миллиардов), что влечет высокие вычислительные затраты и длительное время обработки. PEFT предлагает принципиально иную стратегию: вместо глобальной перенастройки он предполагает модификацию лишь малой доли параметров базовой модели за счет внедрения специализированных компонентов — так называемых модулей адаптации (adapter modules). Такие модули вносят минимальное число дополнительных параметров и эффективно обучаются даже на ограниченных наборах данных.

Принцип действия PEFT базируется на трех положениях:

1. Базовая LLM сохраняется в неизменном виде.
2. К ней добавляется компактная модульная подсистема (например, adapter layer) с небольшим числом параметров.
3. Адаптация к конкретной задаче затрагивает исключительно эти дополнительные параметры.

Это существенно сокращает время дообучения и снижает нагрузку на вычислительные ресурсы, позволяя крупным моделям быстро адаптироваться к узким задачам без переобучения всей архитектуры.

Среди распространенных реализаций PEFT выделяют:

1. Adapter Modules – компактные слои, встраиваемые в каждую позицию трансформера и хранящие минимум параметров для целевой адаптации;

2. LoRA (Low-Rank Adaptation) – метод, добавляющий к сети низкоранговые матрицы, которые обучаются исключительно под конкретную задачу;

3. Prompt Learning – подход, вовсе не модифицирующий параметры модели, а генерирующий специальные подсказки (prompts) для ввода дополнительной информации при решении задачи [3].

Использование mem-векторов

Модели трансформеров традиционно страдают от недостатка долговременной памяти, так как информация, обработанная на ранних стадиях, теряется в процессе передачи сигнала через слои. Чтобы решить эту проблему, были предложены mem-векторы – специальный вид вспомогательных представлений, предназначенных для накопления и последующего повторного использования важной информации.

Mem-вектор (memory vector) – это дополнительное векторное представление, которое интегрируется в архитектуру трансформера и служит своего

рода «банком памяти». Когда модель проходит этап обучения или обработки данных, она записывает особо ценные фрагменты информации в эти mem-векторы, к которым затем может обратиться позднее при выполнении последующих шагов.

Пример: в диалоговом боте, использующем mem-векторы, каждое сообщение собеседника может ассоциироваться с уникальным mem-вектором, который впоследствии используется для поддержания непрерывности беседы и лучшего понимания контекста.

Преимущества использования mem-векторов

1. Долговременная память: позволяет трансформеру лучше удерживать важную информацию на протяжении долгих периодов времени, улучшая его способность понимать контекст и действовать адекватно ситуации.

2. Расширяемость: модель может поддерживать любое количество mem-векторов, увеличивая общую емкость памяти без изменения базовой архитектуры.

3. Повышенная точность: поскольку mem-векторы сохраняют ключевые моменты контекста, модель принимает более обоснованные решения при классификации, генерации текста или других задачах [4].

Практические приложения и результаты

Эффективность и производительность современных архитектур трансформеров позволили достичь значительных успехов в различных областях. Они обеспечивают высокую точность и гибкость, одновременно уменьшая потребности в ресурсах и энергии, что делает их доступными даже для небольших устройств.

Методики параметрического эффективного дообучения (например, LoRA) упростили процесс обновления больших языковых моделей, сделав его быстрее и дешевле. Добавление минимальных дополнительных параметров снижает затраты на обучение, сохраняя качество результата.

Разработка mem-векторов позволила расширить память трансформеров, эффективно сохраняя важную информацию для решения сложных задач. Такие подходы помогают справляться с большими объемами данных и обеспечивать высокое качество работы.

Примером успешного внедрения трансформеров в медицинскую сферу стала система AlphaFold, разработанная DeepMind. Она обеспечивает точное прогнозирование структуры белков, что способствует ускоренному поиску новых лекарственных препаратов.

Образование также выигрывает от применения трансформеров. Автоматизированные системы оценки успеваемости и персонализации учебных планов способствуют улучшению качества образования, помогая учителям эффективнее организовывать учебный процесс.

Наконец, креативные индустрии получили мощный импульс благодаря новым алгоритмам на основе трансформеров. Искусственный интеллект способен создавать уникальные художественные произведения, тексты и музыкальное содержание, стимулируя развитие творческого потенциала [5–7].

Заключение

Исследование показало, что эволюция архитектур трансформеров прошла путь от простых механизмов самовнимания до сложных sparse-моделей, оптимизированных для снижения вычислительных затрат и повышения эффективности.

Основные выводы исследования

1. Развитие sparse-моделей. Современные sparse-модели позволяют обрабатывать длинные последовательности данных с минимальными ресурсами, обеспечивая высокую производительность и энергоэффективность.

2. Эффективность fine-tuning. Параметрически эффективные методы дообучения (PEFT) позволили снизить затраты на обучение и сделать крупные языковые модели (LLMs) доступными для широкого круга пользователей.

3. Расширение возможностей памяти. Использование mem-векторов и внешних хранилищ знаний позволило трансформерам сохранять и повторно использовать важную информацию, улучшив их способность понимать контекст и решать комплексные задачи.

Эти достижения расширяют область применения трансформеров, делая их полезными инструментами в медицине, биологии, образовании и искусстве. Дальнейшие исследования будут направлены на совершенствование методов оптимизации, расширение функциональных возможностей и создание более универсальных моделей, способных справляться с разнообразными задачами.

ЛИТЕРАТУРА / REFERENCES

1. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., et al. Attention is all you need. arXiv [Preprint]. 2017. Available from: https://arxiv.org/abs/1706.03762 (accessed 12 Jun 2026).

2. Child R., Gray S., Radford A., Sutskever I. Generating long sequences with sparse transformers. arXiv [Preprint]. 2019. Available from: https://arxiv.org/abs/1904.10509 (accessed 12 Jun 2026).

3. Hu E.J., Shen Y., Wallis P., Allen-Zhu Z., Li Y., Wang S., et al. LoRA: Low-rank adaptation of large language models. arXiv [Preprint]. 2021. Available from: https://arxiv.org/abs/2106.09685 (accessed 12 Jun 2026).

4. Bulatov A., Kuratov Y., Burtsev M.S. Recurrent memory transformer. arXiv [Preprint]. 2022. Available from: https://arxiv.org/abs/2207.06881 (accessed 12 Jun 2026).

5. Ramesh A., Pavlov M., Goh G., Scott G., Voss C., Radford A., et al. Zero-shot text-to-image generation. arXiv [Preprint]. 2021. Available from: https://arxiv.org/abs/2102.12092 (accessed 12 Jun 2026).

6. Dhariwal P., Jun H., Payne C., Kim J.W., Radford A., Sutskever I. Jukebox: A Generative model for music. arXiv [Preprint]. 2020. Available from: https://arxiv.org/abs/2005.00341 (accessed 12 Jun 2026).

7. Brown T.B., Mann B., Ryder N., Subbiah M., Kaplan J., Dhariwal P., et al. Language models are few-shot learners. arXiv [Preprint]. 2020. Available from: https://arxiv.org/abs/2005.14165 (accessed 12 Jun 2026).

V. O. GULYAEV, O. I. ZAKHAROVA

EVOLUTION OF TRANSFORMER ARCHITECTURES: FROM SELF-ATTENTION TO EFFICIENT SPARSE MODELS

Volga State University of Telecommunications and Informatics
Samara, Russian Federation

Abstract. *The article is devoted to the study of new architectures of artificial neural networks aimed at improving classical transformers by reducing computational costs and increasing efficiency. The main stages of the evolution of transformers are considered, starting with the introduction of the mechanism of self-attention and ending with modern sparse models. Particular attention is paid to optimization techniques such as parametrically efficient fine-tuning (PEFT), adaptive layers (adapters), as well as the use of external knowledge repositories and memory vectors to expand the long-term memory of models. The article concludes with a discussion of current achievements, limitations, and promising areas for future research in this area.*

Keywords**:** *transformers, self-attention mechanism, sparse models, PEFT, memory vectors*

Гуляев Владислав Олегович
Поволжский государственный университет телекоммуникаций и информатики, г. Самара, Российская Федерация. Магистрант.
Vladislav O. Gulyaev
Volga State University of Telecommunications and Informatics, Samara, Russian Federation. Master's student.
E-mail: vladislavgulaev03@gmail.com

Захарова Оксана Игоревна
Поволжский государственный университет телекоммуникаций и информатики, г. Самара, Российская Федерация. Заместитель заведующего Научно-исследовательской лабораторией искусственного интеллекта.
Oksana I. Zakharova
Volga State University of Telecommunications and Informatics, Samara, Russian Federation. Deputy Head of the Research Laboratory of Artificial Intelligence.
E-mail: o.zaharova@psuti.ru