ORES
Научно-практический электронный журнал
Оригинальные исследования (ОРИС), *ISSN 2222-0402*
www.ores.su
original-research.ru

АНАЛИЗ И ДЕКОМПОЗИЦИЯ АРХИТЕКТУРЫ ТРАНСФОРМЕРА В РАМКАХ РАЗВИТИЯ ГЕНЕРАТИВНО ПАРАМЕТРИЧЕСКИХ МОДЕЛЕЙ НЕЙРОННЫХ СЕТЕЙ

Павловский Денис Валерьевич,
студент, магистрант
1 курс, факультет “Отдел магистратуры”, кафедра «Медиатехнологии»
Донской государственный технический университет
Россия, г. Ростов-на-Дону

Кудинов Никита Георгиевич,
руководитель Центра R&D, Технопарк Магика,
генеральный директор, ООО «ЗРЕНИЕ 2.0» (резидент Фонда «Сколково»)
аспирант по направлению «Медиакоммуникация и журналистика»
Донской государственный технический университет
Россия, г. Ростов-на-Дону

Аннотация

В рамках данной работы проанализированы основные компоненты и механизмы архитектуры нейронных сетей по типу архитектуры трансформера, приведена общая схема архитектуры. А также рассмотрены модификации моделей генеративно параметрических трансформерах для решения задач связанных с понимание языка и генерацией текстов. Приведены ключевые различия в архитектуре существующих больших языковых моделей (LLM), и их вклад в развитие архитектуры. Также подобные модели имеют широкий спектр применений, включая генерацию контента, аудио-видео обработку, искусственный дизайн и другие области.

Ключевые слова: искусственный интеллект, нейронные сети, генеративно параметрические сети

ANALYSIS AND DECOMPOSITION OF TRANSFORMER ARCHITECTURE IN THE FRAMEWORK OF THE DEVELOPMENT OF GENERATIVE PARAMETRIC MODELS OF NEURAL NETWORKS

Denis V. Pavlovskiy,
student, master's student
1st year, Faculty "Master's Degree Division", Department "Media Technologies"
Don State Technical University
Russia, Rostov-on-Don

Nikita G. Kudinov,
Head of R&D Center, Technopark Magika,
General Director, LLC "Sight 2.0" (Skolkovo Foundation resident)

Postgraduate student in the field of Media Communication and Journalism
Don State Technical University
Russia, Rostov-on-Don

ABSTRACT

Within the framework of this work, the main components and mechanisms of the architecture of neural networks according to the type of transformer architecture are analyzed, and the general architecture scheme is given. Modifications of generative parametric transformers models for solving problems related to language understanding and text generation are also considered. The key differences in the architecture of existing large language models (LLM) and their contribution to the development of architecture are presented. Also, such models have a wide range of applications, including content generation, audio-video processing, artificial design and other areas.

Keywords: artificial intelligence, neural networks, generative parametric networks

Введение. Трансформаторная архитектура — это один из видов архитектуры нейронных сетей, который широко используется для решения задач, связанных с обработкой последовательных данных, например: текста или временных рядов. В настоящее время нейронные сети, построенные с помощью этого подхода, используются при проектировании генеративно параметрических сетей для задач, связанных с обработкой различного медиаконтента. Первоначально разработанная для обработки текста, архитектура трансформера состоит из комплексных уровней внимания и уровней кодера и/или декодера, которые вместе позволяют последовательно создавать экземпляры долгосрочных зависимостей. Нейронные сети, построенные как генеративно параметрические трансформеры, используются чтобы учитывать сложные контекстные наборы данных, что делает генеративно-параметрические трансформеры эффективным инструментом обработки информации и создания контента.

Основные концепции архитектуры трансформера.

Система трансформера состоит из двух основных компонентов: энкодера и декодера. Каждый из них состоит из нескольких компонентов, называемых блоками-трансформерами. Также каждый из трансформерных блоков содержит механизм внимания (attention mechanism), благодаря которому модель может обращать внимание на различные части входных данных и интерпритировать их в зависимости от их значимости в рамках контекста.

Энкодер трансформера, в качестве входного занчения, принемает последовательность символов и преобразует ее в векторное представление. Он состоит из нескольких блоков, где каждый блок содержит две основные операции: механизм внимания (attention) и полносвязная сеть с пропусками.

Декодер трансформера принимает на вход векторное представление, полученное от энкодера, и генерирует последовательность символов. Он также состоит из нескольких блоков внимания и полносвязную сеть с пропусками, но имеет дополнительный механизм внимания, называемый самовнимание (self-attention),также называемый вниманием с маской и схожий в механизме своей работы с меанизмом внимания (attention).

Самовнимание позволяет моделировать зависимости между элементами входной последовательности внутри декодера.

Механизм внимания. В модели трансформера механизм внимания реализуется через множество ключевых точек, каждая из которых вычисляет веса для каждого токена в последовательности на основе его отношений с другими токенами. Затем взвешенные значения токенов объединяются, чтобы получить новое представление последовательности. Этот процесс повторяется несколько раз, позволяя модели учитывать различные аспекты входных данных на разных уровнях абстракции.[1].

В общем случае архитектура трансформера представлена на рисунке1.

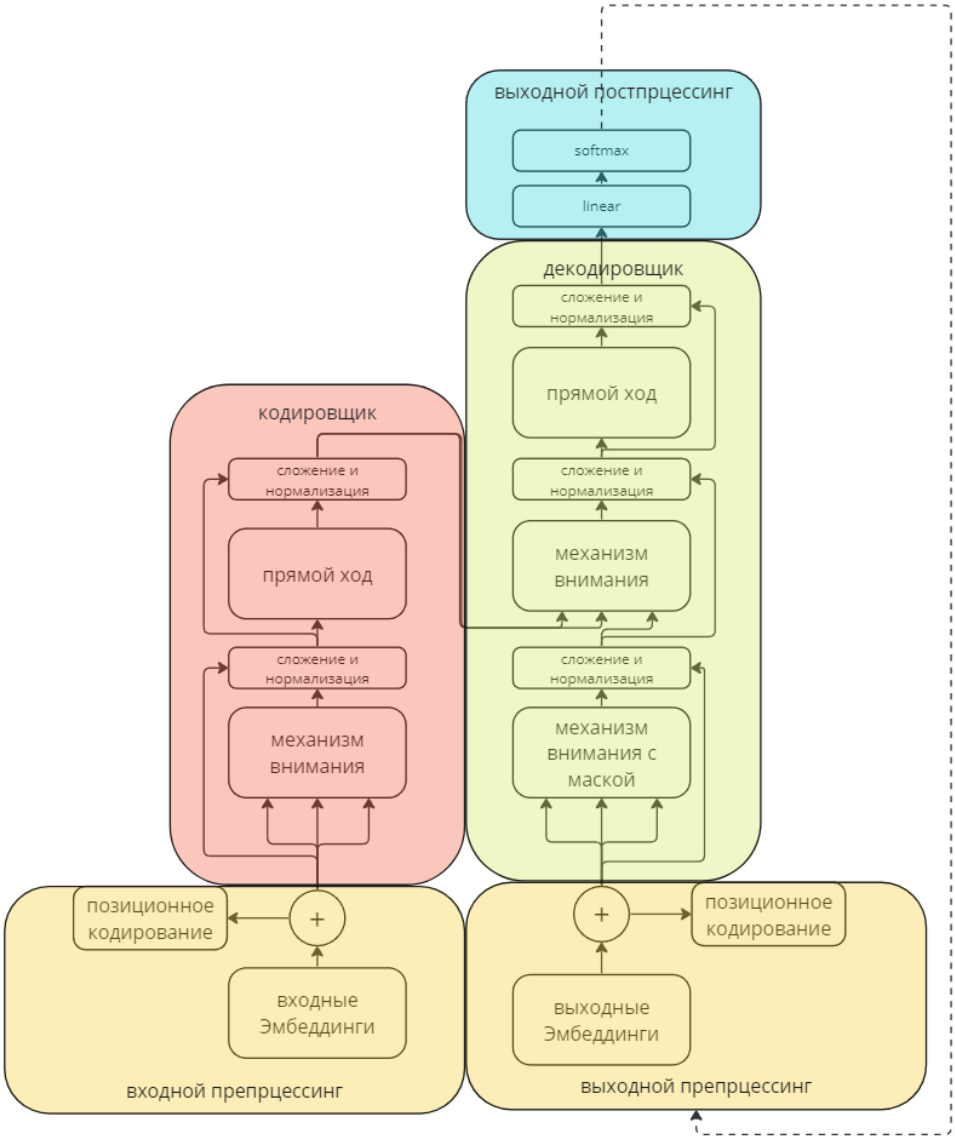

Рисунок 1 – Общая структура архитектуры трансформера

Однако, важно заметить, что возможно и построение нейронной сети и с использованием слоев только энкодера или декодера.

Двунаправленные трансформеры. Стандартные трансформеры анализируют последовательности слов или других элементов входных данных только в одном направлении, что может привести к потере контекста в противоположном направлении. Это может стать проблемой в том случае, если последовательность важна. В работе "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" командой исследователей из Google AI Language[2]. была представлена модификация модели трансформера, которая способная учитывать контекст, находящийся как слева, так и справа от текущего элемента в позиции. Эта модификация была названа двунаправленным

трансформером Это архитектурное решение решает проблему потери контекста с одной из сторон за счет использование двух слоев внимания: один для прямого последовательного анализа и один для обратного, впоследствии объединяя результаты обеих систем для расчета соответствующей контекстуальной информации с учетом обеих сторон Таким образом двунаправленные трансформеры могут учитывать более широкий контекст при анализе последовательностей. Важно отметить, что такая модификация увеличивает общую вычислительную сложность системы, однако в дальнейшем производительность может быть улучшена за счет применения особых методов, таких как например квантизация весов и слияние групп слоев.

Применение генеративно параметрических трансформеров. Большое развитие архитектура трансформера в генеративно параметрических задачах получила в рамках построения больших языковых моделей (LLM). Различные компании использовали как подход с использованием только слоев энкодера, так и смешанный подход и подход с использованием только декодера.

Общая схема развития архитектуры генеративно параметрических трансформеров, в контексте больших языковых моделей представлена на рисунке 2.

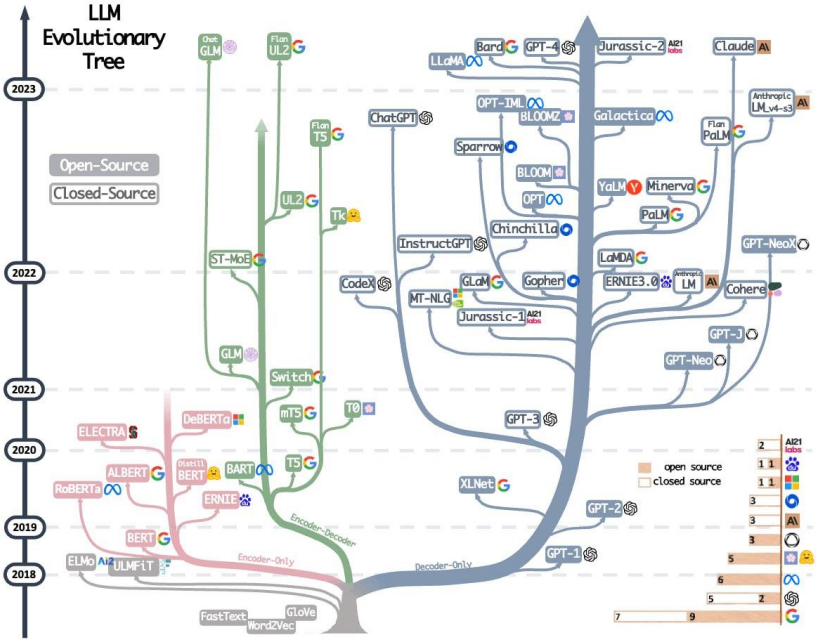

Рисунок 2 – Развитие больших языковых моделей (LLM)

Перспективы и развитие. Перспективы развития ИИ на базе архитектуры трансформера включают следующие аспекты:

Улучшение качества моделей: генеративно параметрические трансформеры демонстрируют высокую точность в задачах, связанных с обработкой естественного языка, таких как машинный перевод, анализ тональности, вопросно-ответные системы и др. [3]. Однако в процессах, связанных с пониманием отельных лексем или использованием контекста, связанного с распознаванием символов, остается возможность и необходимость улучшения точности модели.

Расширение области применения: Архитектура трансформера также может быть применена в других областях, помимо обработки естественного языка и создания тестового контента. Например, уже сейчас, она может быть использована для анализа временных рядов [4], обработки изображений или даже для моделирования графовых данных. Но остаются задачи машинного обучения, где может быть достигнут прогресс за счет адаптации модели трансформера для учета контекстуальных зависимостей.

Оптимизация и ускорение: Трансформеры, особенно при работе с большими объемами данных, могут быть вычислительно требовательными. В связи с этим одним из актуальных направлений развития является оптимизация и ускорение работы сетей на базе трансформера, чтобы они стали более эффективными в использовании ресурсов и могли быть применены для решения задач в реальном времени [5].

Заключение. Таким образом, ключевыми факторами в рамках развития генеративно параметрических трансформеров является механизм внимания, позволяющий нейронной сети учитывать контекст как при генерации выходных значений, так и при анализе входящего запроса и/или последовательности данных. Также важную роль в применение трансформеров имеет архитектура двунаправленных трансформеров, а дельнейшее развитие данной архитектуры заключается в оптимизации скорости вычисления и обучения. Важно отметить, что в настоящее время нет однозначной закономерности между качеством работы обученной сети и соотношением слоев кодера/декодера.

Список литературы:

1. Dai Z. et al. Transformer-xl: Attentive language models beyond a fixed-length context //arXiv preprint arXiv:1901.02860. – 2019.
2. Devlin J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding //arXiv preprint arXiv:1810.04805. – 2018.
3. Анализ возможностей языковых моделей BERT и chatgpt для понимания естественного языка / Р. И. Ким, А. В. Андреев, А. Г. Базанова [и др.] // Информация и образование: границы коммуникаций. – 2023. – № 15(23). – С. 298-300. – DOI 10.59131/2411-9814_2023_15(23)_298. – EDN GSZVKW.
4. Mazzia V. et al. Action Transformer: A self-attention model for short-time pose-based human action recognition //Pattern Recognition. – 2022. – Т. 124. – С. 108487.
5. Tay Y. et al. Lightweight and efficient neural natural language processing with quaternion networks //arXiv preprint arXiv:1906.04393. – 2019.

References:

1. Dai Z. et al. Transformer-xl: Attentive language models beyond a fixed-length context //arXiv preprint arXiv:1901.02860. – 2019.
2. Devlin J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding //arXiv preprint arXiv:1810.04805. – 2018.
3. Analysis of the capabilities of the BERT and chatgpt language models for understanding natural language / R. I. Kim, A.V. Andreev, A. G. Bazanova [et al.] // Information and education: boundaries of communications. – 2023. – № 15(23). – C. 298-300. – DOI 10.59131/2411-9814_2023_15(23)_298. – EDN GSZVKW.
4. Mazzia V. et al. Action Transformer: A self-attention model for short-time pose-based human action recognition //Pattern Recognition. – 2022. – Т. 124. – С. 108487.
5. Tay Y. et al. Lightweight and efficient neural natural language processing with quaternion networks //arXiv preprint arXiv:1906.04393. – 2019.