Архитектура динамической децентрализованной большой языковой модели на основе блокчейна

Н. В. Козгунов, М. Халаши
Санкт-Петербургский государственный университет
st090866@student.spbu.ru

В. Д. Олисеенко
Санкт-Петербургский Федеральный исследовательский центр РАН
vdo@dscs.pro

Т. В. Тулупьева
Санкт-Петербургский Федеральный исследовательский центр РАН
Северо-Западный институт управления РАНХиГС
tvt@dscs.pro

Аннотация. **В работе представлена новая архитектура большой языковой модели на принципе блокчейн. В основе архитектуры лежит концепция федеративного обучения через сеть узлов, призванного улучшить процесс обучения и использования LLM с применением вычислительной мощности и данными для обучения, которые предоставляют сами пользователи блокчейна. Представленная архитектура направлена на устранение проблем централизации, цензурирования и предвзятости в существующих LLM, делая доступнее их использование пользователям.**

Ключевые слова: дифференциальная конфиденциальность; обработка естественного языка; децентрализованное федеративное обучение; блокчейн; Web 3.0

I. Введение

Начало 2010-х годов ознаменовалось значительным прогрессом в области обработки естественного языка, начиная с появления Word2Vec [1], ELMo [2] и Transformer [3], что привело к созданию больших языковых моделей (LLM), таких как семейства BERT [4] и GPT [5]. Развитие области обработки естественного языка помогло повысить понимание синтаксиса и семантики моделями машинного обучения [4–5].

Одновременно с этим появление концепции блокчейна, начавшегося с Bitcoin [6] и Ethereum [7], заложило основу для создания нового класса распределенных вычислений и концепции Web 3.0.

Совместное использование блокчейна и LLM может объединить преимущества технологий децентрализации блокчейна, повысив конфиденциальность, создав условия для оптимизации использования ресурсов, повысив защиту от сбоев централизованного сервера, а также снизив затраты, связанные с дообучением и использованием LLM [6–7, 14, 17, 19–20]. Стоит отметить, что такие теоретические решения находятся на начальном этапе развития, что подтверждается в некоторых публикациях [14, 19, 21].

Таким образом, цель данной работы заключается в разработке и описании архитектуры большой языковой модели на основе технологии блокчейн.

Исследование выполнено в рамках проекта по государственному заданию СПб ФИЦ РАН № FFZF-2022-0003

II. Обзор литературы

Системы федеративного машинного обучения за счёт использования различных устройств для хранения данных и обучения моделей имеют существенные преимущества над централизованными системами машинного обучения благодаря обеспечению конфиденциальности данных и оптимизации процессов масштабируемости [8]. Однако они подвержены определенным недостаткам, включая коммуникационные издержки и риски единой точки сбоя из-за зависимости от локальных серверов [9]. В качестве альтернативы для решения указанных недостатков может выступать система децентрализованного федеративного обучения, которая использует механизмы гомоморфного шифрования для возможности обеспечения безопасного и эффективного обучения [10]. Также система предлагает структуру, основанную на ориентированных ациклических графах, для обеспечения надежности децентрализованных операций, стимулируя более широкое участие и обеспечивая справедливые выплаты участникам в соответствии с вкладом их вычислительных мощностей в обучение и вывод модели [11].

Потенциал интеграции федеративного обучения и децентрализованных методов с LLM значителен [12–16]. Исследование FusionAI демонстрирует, что потребительские графические процессоры могут эффективно использоваться для обучения и LLM, предлагая экономически выгодную альтернативу относительно существующих языковых моделей за счет сокращения расходов на мощности графических процессоров для дообучение и хранения существующей модели, расширяя возможности вывода (*англ. inference*) LLM на мобильные устройства [12, 14], при этом обеспечивая высокопроизводительный генеративный вывод в условиях ограниченных ресурсов [15]. Исследования показывают, что улучшение функциональности LLM на индивидуальных узлах, особенно в распределенных конфигурациях, способствует повышению общей эффективности системы [14–16].

Интеграция технологий консенсусов Proof-of-Work (*PoW*), Proof-of-Work (*PoS*), Proof-of-Engagement (*PoE*) [18] и Proof-of-Reputation (*PoR*) [19] в LLM представляет

собой перспективное направление развития области языковых моделей. Современный подход к децентрализации алгоритмов машинного обучения и данных через использование консенсусов направлен на усиление безопасности, приватности и коллаборативности. В контексте этих разработок сочетание больших языковых моделей и блокчейн-технологий открывает новые возможности, совмещая контекстуальное понимание и творческие способности языковых моделей с безопасностью и прозрачностью блокчейна, предлагая эффективное синергетическое взаимодействие [20].

Несмотря на ряд работ, посвященных децентрализованным большим языковым моделям [12–14], в них не использовались технологии блокчейна, которые могли бы усовершенствовать дообучение модели, защиту от сбоев и хранение информации распределенно, благодаря надежным механизмам стимулирования (криптовалютами) и протоколам консенсуса. Такая интеграция может значительно расширить процессы обучения и вывода, сделав их более эффективными и быстрыми. Кроме того, блокчейн может способствовать динамическому дообучению LLM посредством обновления версий на основе сетевого консенсуса, где версия модели при подключении к сети обновляется до последней. Хотя в некоторых исследованиях [16, 21–22] интеграция уже изучалась, в основном они были посвящены использованию LLM в блокчейне специально для улучшения децентрализованных приложений. В работе предлагается новая архитектура, которая способствует динамическому дообучению и локальным выводам LLM с системой поощрений, основанных на консенсусах. Эта среда поощряет участие пользователей, предлагая стимулы (криптовалюту) для предоставления данных, памяти и вычислительных ресурсов, используя механизмы консенсуса для обеспечения целостности сети, динамического обновления моделей и облегчения версионирования моделей.

III. Архитектура динамической децентрализованной большой языковой модели

A. *Обзор*

Веса в предлагаемой архитектуре первоначально устанавливаются с помощью методов трансферного обучения, которые затем распространяются по сети в одноранговом режиме. За этим шагом следует локальное обновление модели с использованием подходов федеративного обучения для уточнения модели на децентрализованных узлах. Затем локальные модели объединяются в новую, расширенную версию LLM, которая снова распространяется по сети. В основе всего процесса лежит блокчейн в качестве механизма для обновления и контроля версий LLM, включающий криптовалютную схему поощрения для мотивации участников сети.

Упрощенная схема предлагаемой архитектуры представлена на рис. 1.

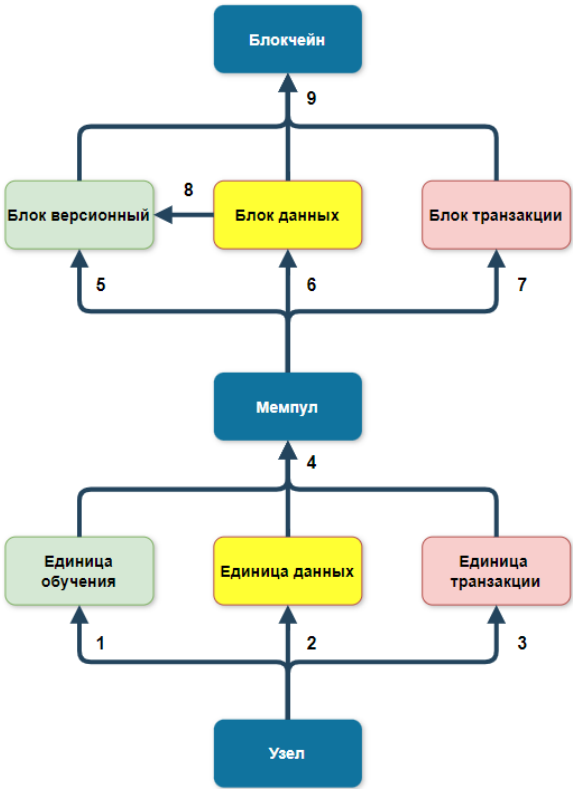

Рис. 1. Архитектура динамической децентрализованной большой языковой модели на основе блокчейна

B. *Компоненты ахритектура*

Предлагаемая архитектура модели состоит из следующих элементов, представленных на рис. 1:

- **Узел** (*англ. Node*) – конечные устройства сети, предоставляющие вычислительные ресурсы и данные.
- **Единица обучения** (*англ. Learning unit*) – обновленные веса модели, полученные узлами в результате локального обучения на своих наборах данных.
- **Единица данных** (*англ. Data unit*) – фрагмент набора данных, который узлы передают по сети.
- **Единица транзакции** (*англ. Transaction unit*) – транзакции монет между узлами в сети.
- **Мемпул** (*англ. Mempool*) – место для хранения единиц данных, результатов единиц обучения, записи единиц транзакции перед их обновлением в новые блоки.
- **Блок транзакции** (*англ. Transaction block*) – список из нескольких транзакций, выбранных из пула памяти.
- **Блок данных** (*англ. Data block*) – блок, включающий транзакцию coinbase для получения вознаграждения и компиляцию блоков данных из Мемпула, формирующих тестовый набор данных.
- **Версионный Блок** (*англ. Version block*) – блок, содержащий транзакцию coinbase за вознаграждение и агрегированные веса, формирующие глобальную модель, полученную из комбинации единиц обучения Мемпула.

C. Описание работы

Работа модели основана на представленной архитектуре (рис.1) начинается с установки начальных весов на выбранном узле путем трансферного обучения, используя веса из предварительно обученной модели, а не по принципу случайного выбора. Затем эти веса распространяются по одноранговой сети (P2P), позволяя узлам индивидуально обучать модель на своих данных. После обучения узлы обновляют веса модели и переводят эти обновления в блок обучения (рис. 1, 1). Этот блок защищается с помощью шифрования с закрытым ключом [22], создавая уникальную цифровую подпись, и впоследствии отправляется в Мемпул [23]. Для поддержки последующего комбинирования моделей применяется гомоморфное шифрование [22].

Одновременно узлы совершают криптовалютные транзакции в блокчейне, и эти транзакции (рис. 1, 3) также направляются в Мемпул (рис. 1, 4). Более того, некоторые узлы предоставляют тестовые наборы данных, шифруя и подписывая единицы данных в процессе (рис. 1, 2), схожем с процессом обучения образом, перед тем как отправить их в Мемпул (рис. 1, 4), где гомоморфное шифрование позволяет выполнять последующие операции.

После этого этапа механизм PoS выбирает узлы на основе размера их криптовалютных аткивов для создания новых блоков, добавляя определенную степень случайности для обеспечения разнообразного выбора создателей блоков. Эти узлы извлекают, проверяют и собирают транзакции из Мемпул в новый блок, который включает специальную транзакцию coinbase [23] в качестве вознаграждения для создателя блока (рис. 1, 7). Затем этот блок распространяется по сети для проверки целостности. Если большинство узлов подтверждают транзакции, блок интегрируется в блокчейн (рис. 1, 9); в противном случае недействительные транзакции приводят к потере доли создателя блока, что стимулирует тщательную проверку.

Система блокчейна устанавливает ограничение на частоту блоков транзакций, требуя добавления блока данных (рис. 1, 5) и версионного блока (рис. 1, 6) перед любыми последующими блоками транзакций. Эта процедура использует комбинацию механизмов PoS, Proof-of-Time (PoT) и PoW. Для генерации блока данных PoS определяет выбор узла на основе заданных узлами ставок, при назначении случайного вызова (номера), который обрабатывается с помощью функции верифицируемой задержки (VDF). Узлы выбирают и улучшают блоки данных из Мемпула (рис. 1, 5) для повышения качества набора данных. Создатель набора данных наивысшего качества формирует новый блок данных, который затем аутентифицируется сетью с помощью выходных данных VDF. Создатели проверенных блоков данных получают вознаграждение через транзакцию coinbase, отражающее их вклад в качество набора данных. Узлы, искажающие качество данных или манипулирующие VDF, рискуют лишиться вознаграждения.

В соответствии с установленной политикой частоты блоков, блокчейн останавливает включение блоков, не относящихся к транзакциям, до тех пор, пока не будет добавлен блок новой версии (рис. 1, 8). Используя механизм PoS, узлы выбираются для создания нового версионного блока. Этот процесс отбора, подобный тому, что используется при создании блока данных, направлен на объединение обучаемых блоков для повышения производительности модели. Уникальным аспектом этого этапа является разделение тестового набора данных, полученного из предыдущего блока данных, на различные партии (*англ. batches*). Затем эти партии распределяются между выбранными узлами. Выдача вызова вместе с этими разнообразными партиями данных побуждает узлы применять усреднение с помощью федеративного усреднения [24] или аналогичными методами для уточнения агрегированной модели на основе полученных данных.

После завершения процесса верификации с помощью функции проверяемой задержки узла, достигшему наивысшей производительности по всему набору тестовых данных, поручается создание блока новой версии. Далее блок проходит верификацию во всей сети. Успешный процесс проверки не только подтверждает правильность версионного блока, но и вызывает вознаграждение как для создателя новой версии блока, так и для узлов, предоставивших единицы обучения, использованные при его разработке. И наоборот, любой узел, уличенный в злоупотреблениях во время этого процесса, рискует потерять свою долю, что обеспечивает целостность и поощряет честное участие в работе блокчейна.

После успешной агрегации и проверки обновленной модели, заключенной в версионном блоке блокчейна, узлы получают возможность использовать эту улучшенную модель для решения задач вывода. У них есть возможность либо использовать свои локальные вычислительные ресурсы, что может привести к менее мощным выводам по сравнению с использованием коллективных вычислительных возможностей всей сети, либо получить доступ к более широкой вычислительной мощности и ресурсам сети через механизм обмена монетами с другими узлами. Система поощрений подразумевает сбалансированность вклада и выгоды для всей сети: предполагается, что средний узел будет зарабатывать на предоставлении единиц данных и единиц обучения примерно столько же, сколько он потратит на доступ к вычислительным ресурсам для выводов. Этот подход направлен на минимизацию стоимости использования системы для среднего узла. В отличие от этого, крупные узлы, требующие более обширных выводов и высокой точности, могут понести дополнительные расходы на заимствование необходимых вычислительных мощностей из сети.

IV. Обсуждение

Представленная архитектура сталкивается с рядом проблем, главная из которых – получение вредоносных данных, влияющих на результаты моделирования. Также использование неидентичных и неидентично распределенные данные (non-IID) и перекос данных могут повлиять на эффективность дообучения модели. Для решения проблемы качества наборов данных, необходимых для создания блоков данных и версионных блоков, может быть использован комбинированный

подход, сочетающий различные метрики качества текста, например, статистические показатели, такие как показатели качества содержания; индекс Ганнинга-Фога; показатели разнообразия и сложности; мера текстового лексического разнообразия; а также показатели новизны и избыточности, например, коэффициент Жаккара. Оценка эффективности модели может включать в себя совокупность таких метрик, как accuracy, F1 Score, BLEU Score и ROUGE Score. Эти оценки должны быть вычислительно эффективными, чтобы обеспечить возможность проверки узлами в рамках протокола консенсуса архитектуры.

Еще одна проблема — управление большим объемом данных в блокчейне, сохраняя масштабируемость. Решения могут включать в себя стратегии вне цепочки (*англ. off-chain*), такие, как децентрализованные сети хранения (DSN) или якорение данных, а также шардинг данных и решения для масштабирования второго уровня. Кроме того, задержка, часто встречающаяся в сетях блокчейн, создает проблемы и в нашей архитектуре. Для ее снижения можно оптимизировать механизмы консенсуса, использовать передовые сетевые протоколы, такие как усовершенствованные протоколы Gossip, применять методы шардинга и исследовать решения второго уровня. Наконец, сложность реализации этой архитектуры и протоколов, обеспечивающих целостность блокчейна и, сохраняющих конфиденциальность и доступность системы с использованием локальной памяти со снижением предвзятости, весьма значительна. Успешная реализация требует пристального внимания к деталям и следования принципам архитектуры для решения проблем перенаправления вычислительных мощностей с вычисления криптографических задач блокчейна на процессы использования и дообучения LLM.

III. Заключение

В данной статье описывается теоретическая архитектура, объединяющая технологии LLM и блокчейна для создания масштабируемого и распределенного искусственного интеллекта, одновременно решая проблемы конфиденциальности и централизованного контроля. Потенциал такого слияния огромен, оно может повысить эффективность LLM и расширить доступность для пользователей.

Тем не менее, архитектура сталкивается с такими проблемами, как обеспечение конфиденциальности данных, целостности модели и соответствия нормативным требованиям. Ключевые шаги для дальнейшего развития архитектуры включают детальную доработку архитектуры, протоколов консенсуса и разработку пилота программной реализации для тестирования.

В дальнейших работах также предстоит синтезировать эффективные метрики и алгоритмы для оценки данных и производительности, а также решить проблему загруженности блокчейна данными.

Список литературы

[1] Caselles-Dupré H., Lesaint F., Royo-Letelier J. Word2vec applied to Recommendation: Hyperparameters Matter // arXiv. Aug 2018. URL: https://arxiv.org/abs/1804.04212 (дата обращения: 08.01.2024).

[2] Karampatsis R., Sutton C. SCELMO: SOURCE CODE EMBEDDINGS FROM LANGUAGE MODELS // arXiv. Apr 2020. URL: https://arxiv.org/abs/2004.13214 (дата обращения: 08.01.2024).

[3] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A., Kaiser Ł., Polosukhin I. Attention is all you need In Advances in neural information processing systems // arXiv. Jun 2017. URL: https://arxiv.org/abs/1706.03762 (дата обращения: 08.01.2024).

[4] Alaparthi S., Mishra M. Bidirectional Encoder Representations from Transformers (BERT): A sentiment analysis odyssey // arXiv. Jul 2020. URL: https://arxiv.org/abs/2007.01127 (дата обращения: 08.01.2024).

[5] Yenduri G., Ramalingam M., Chemmalar G., Supriya Y., Srivastava, Maddikunta G., Raj G., Jhaveri H., Prabadevi B., Wang W., Vasilakos V., Gadekallu T. GPT (Generative Pre-trained Transformer) – A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions // arXiv. May 2023. URL: URL: https://arxiv.org/abs/2305.10435 (дата обращения: 08.01.2024).

[6] Reid F., Harrigan M. An Analysis of Anonymity in the Bitcoin System // arXiv. May 2012. URL: https://arxiv.org/abs/1107.4524 (дата обращения: 29.01.2024).

[7] Pavloff U., Amoussou-Guenou Y., Tucci-Piergiovanni S. Ethereum Proof-of-Stake and the Probabilistic Bouncing Attack // arXiv. Oct 2022. URL: https://arxiv.org/abs/2210.16070 (дата обращения: 30.01.2024).

[8] Garst S., Dekker J., Reinders M. A comprehensive experimental comparison between federated and centralized learning in bioRxiv // DOI: 10.1101/2023.07.26.550615. Jul 2023. URL: https://www.researchgate.net/publication/372759901_A_comprehensive_experimental_comparison_between_federated_and_centralized_learning (дата обращения: 07.02.2002).

[9] You X., Liu X., Lin X., Cai J., Chen S. Accuracy Degrading: Toward Participation-Fair Federated Learning // IEEE Internet of Things Journal. DOI: 10.1109/JIOT.2023.3238038. Jun 2023. URL: https://ieeexplore.ieee.org/document/10021580 (дата обращения: 07.02.2024).

[10] Bose A., Bai L. A Fully Decentralized Homomorphic Federated Learning Framework // IEEE 20th International Conference on Mobile Ad Hoc and Smart Systems (MASS). DOI: 10.1109/MASS58611.2023.00029. Sep 2023. URL: https://ieeexplore.ieee.org/document/10298305 (дата обращения: 14.02.2024).

[11] Yu G., Wang X., Sun C., Wang Q., Yu P., Ni W., Liu R., Xu X. IronForge: An Open, Secure, Fair, Decentralized Federated Learning // arXiv. Jan 2023. URL: https://arxiv.org/abs/2301.04006 (дата обращения: 14.02.2024).

[12] Tang Z., Wang Y., He X., Zhang L., Pan X., Wang Q., Zeng R., Zhao K., Shi S., He B., Chu X. FusionAI: Decentralized Training and Deploying LLMs with Massive Consumer-Level GPUs in Symposium on Large Language Models (LLM 2023) with IJCAI 2023, Macao, China // arXiv. Aug 2023. URL: https://arxiv.org/abs/2309.01172 (дата обращения: 28.02.2024).

[13] Chen C., Feng X., Zhou J., Yin J., Zheng X. Federated Large Language Model: A Position Paper, Zhejiang University, Hangzhou, China // arXiv. 18 Jul 2023. URL: https://arxiv.org/abs/2307.08925 (дата обращения: 29.02.2024).

[14] Zhao J., Song Y., Liu S., Harris I., Jyothi S. LinguaLinked: A Distributed Large Language Model // arXiv. Dec 2023. URL: https://arxiv.org/abs/2312.00388 (дата обращения: 29.02.2024).

[15] Sheng Y., Zheng L., Yuan B., Li Z., Ryabinin M., Fu D., Xie Z., Chen B., Barrett C., Gonzalez J., Liang P., Re´ C., Stoica I., Zhang C. FlexGen: High-Throughput Generative Inference of Large Language Models with a Single GPU // arXiv. Jun 2023. URL: https://arxiv.org/abs/2303.06865 (дата обращения: 06.03.2024).

[16] Alizadeh K., Mirzadeh I., Belenko D., Khatamifard S., Cho M., Mundo C., Rastegari M., Farajtabar M. LLM in a flash: Efficient Large Language Model Inference with Limited Memory // arXiv. Jan 2024. URL: https://arxiv.org/html/2312.11514v2 (дата обращения: 06.03.2024).

[17] Nguyen C., Thai H., Nyato D., Nguyen N., Dutkiewicz E. Proof-of-Stake Consensus Mechanisms for Future Blockchain Networks // IEEE Access. DOI: 10.1109/ACCESS.2019.2925010. Jun 2019. URL: https://ieeexplore.ieee.org/document/8746079?ref=hackernoon.com (дата обращения: 13.03.2024).

[18] Xu Y., Yang X., Zhang J., Zhu J., Sun M., Chen B. Proof of Engagement: A Flexible Blockchain Consensus Mechanism in Wireless Communications and Mobile Computing, Hindawi // DOI: 10.1155/2021/6185910. Aug 2021. URL: https://www.researchgate.net/publication/354037180_Proof_of_Engagement_A_Flexible_Blockchain_Consensus_Mechanism (дата обращения: 13.03.2024).

[19] Do T., Nguyen T., Pham H. Delegated Proof of Reputation: a novel Blockchain consensus, December 2019, arXiv:1912.04065v1 (дата обращения: 14.03.2024).

[20] Liu X. Decentralized Machine Learning on a Blockchain: Case Studies // 2023 IEEE International Conference on Omni-layer Intelligent Systems (COINS). DOI: 10.1109/COINS57856.2023.10189230. Jul 2023. URL: https://ieeexplore.ieee.org/document/10189252 (дата обращения: 21.03.2024).

[21] Mboma J., Tshipata O., Kyamakya K., Kambale W. Assessing How Large Language Models Can Be Integrated with or Used for Blockchain Technology: Overview and Illustrative Case Study in 2023 27th International Conference on Circuits, Systems, Communications and Computers (CSCC) // DOI: 10.1109/CSCC58962.2023.00018. Apr 2023. URL: https://www.researchgate.net/publication/376826315_Assessing_How_Large_Language_Models_Can_Be_Integrated_with_or_Used_for_Blockchain_Technology_Overview_and_Illustrative_Case_Study (дата обращения: 21.03.2024).

[22] Liang W., Zhang D., Lei X., Tang M., Li K., Zomaya A. Circuit Copyright Blockchain: Blockchain-Based Homomorphic Encryption for IP Circuit Protection // IEEE Transactions on Emerging Topics in Computing. DOI. 10.1109/TETC.2020.2993032. Jul.-Sep. 2021. (дата обращения: 27.03.2024).

[23] Saad M., Njilla L., Kamhoua C., Kim J., Nyang D., Mohaisen A. Mempool optimization for Defending Against DDoS Attacks in PoW-based Blockchain Systems // 2019 IEEE International Conference on Blockchain and Cryptocurrency (ICBC). DOI: 10.1109/BLOC.2019.8751476. May 2019. URL: https://ieeexplore.ieee.org/abstract/document/8751402 (дата обращения: 28.03.2002).

[24] Plassier V., Durmus A., Moulines É. Federated Averaging Langevin Dynamics: Toward a unified theory and new algorithms // arXiv. Oct 2022. URL: https://arxiv.org/abs/2211.00100v1 (дата обращения: 01.04.2024).