Архитектура динамической децентрализованной большой языковой модели с использование технологии консенсусов

Козгунов Н., СПбГУ, Санкт-Петербург st090866@student.spbu.ru,
Халаши М., СПбГУ, Санкт-Петербург st082966@student.spbu.ru[1]

Аннотация

В данной исследовательской работе представлена новая концепция большой языковой модели с эффективным использованием вычислительных мощностей по средством использования механизмов управления на основе технологии консенсусов и блокчейна, предоставляя стимулы пользователям транзакциями на основе монет, тем самым позволяя возможность масштабируемости. Одновременно, предлагая более защищенный подход для обработки данных и дообучения модели, устраняя проблему единой точки сбоя с помощью децентрализованного федеративного обучения модели, представленная система решает проблемы существующих больших языковых моделей.

Введение

Ранние 2010-е годы ознаменовались значительным прогрессом в области обработки естественного языка, начиная с появления Word2Vec, который заложил основу для последующих инноваций в этой области. Эти ранние достижения привели к разработке ELMo и были ключевыми в создании архитектуры Transformer , вдохновив на развитие больших языковых моделей (LLM), таких как BERT и серия GPT. Эти последовательные инновации совместно повысили понимание синтаксиса и семантики, устанавливая новые стандарты для понимания языка. Одновременно с этим эволюция блокчейна, начавшаяся с Bitcoin и далее развившаяся с Ethereum, преобразила цифровую валюту и интернет в модель, ориентированную на пользователя, Web 3.0. Это развитие, как подробно описывается, расширило его применение до децентрализованных приложений (DApps) и невзаимозаменяемых токенов (NFT). В данной работе рассматривается слияние LLM и технологии блокчейн, предлагая новую архитектуру, которая объединяет лингвистические возможности, выделенные в работах, с децентрализованной природой блокчейна. Это предложение направлено на решение проблем, связанных с конфиденциальностью, предвзятостью и централизацией. Кроме того,

[1]Авторы внесли равный вклад в данную работу.

оно выступает за использование сети блокчейна для сокращения развивающихся затрат, связанных с LLM, стремясь к масштабируемой и устойчивой модели, которая демократизирует доступ к технологиям ИИ и способствует инклюзивному технологическому будущему.

Обзор литературы

Централизованные системы федеративного обучения демонстрируют значительные преимущества, в том числе обеспечение конфиденциальности данных и оптимизацию управленческих процессов, однако они подвержены определенным недостаткам, включая коммуникационные издержки и риски единой точки сбоя, обусловленные зависимостью от централизованных серверов [1]. В качестве альтернативы представляется децентрализованное федеративное обучение, которое внедряет механизмы, основанные на гомоморфном шифровании, для обеспечения безопасного и эффективного обучения через множество источников без необходимости расшифровки, тем самым улучшая устойчивость и производительность системы [2], а также предлагая структуру, основанную на DAG, для обеспечения надежности децентрализованных операций, стимулируя более широкое участие и обеспечивая справедливые стимулы участникам [3]. Потенциал интеграции федеративного обучения и децентрализованных методов с большими языковыми моделями (LLM) значителен. Исследование FusionAI демонстрирует, что потребительские графические процессоры могут эффективно использоваться для обучения и развертывания LLM, предлагая экономически выгодную альтернативу [4]. Федеративные LLM адресуют проблемы дефицита данных и приватности в искусственном интеллекте, применяя коллаборативное обучение, включая процессы предварительного обучения и дообучения, в рамках централизованной модели FL [5]. Предлагаемая архитектура способствует созданию дополнительной оптимизации процесса вывода LLM на мобильные устройства, повышая быстродействия системы [6]. Улучшение функциональности LLM на индивидуальных узлах, особенно в распределенных конфигурациях, способствует повышению общей эффективности системы. Разработка FlexGen обеспечивает высокопроизводительный движок для генеративного вывода в условиях ограниченных ресурсов [7], в то время как адаптация LLM для работы на устройствах с ограниченной оперативной памятью, например, через использование флэш-памяти для хранения данных, значительно ускоряет процесс вывода [8]. Трансформации в блокчейн-технологиях являются ключевыми, с переходом от механизма доказательства выполнения работы (Proof-of-Work, PoW) к механизму доказательства доли (Proof-

of-Stake, PoS), что вносит фундаментальные изменения в механизмы достижения консенсуса через владение долями, имея критическое значение для будущей безопасности и производительности сетей [9]. Механизм доказательства вовлечения (Proof-of-Engagement, PoE) предоставляет гибкую систему консенсуса, акцентируя внимание на активных вкладах узлов, тем самым продвигая принципы равенства [10], в то время как делегированное доказательство репутации (Delegated Proof-of-Reputation) объединяет PoS с системой репутации для повышения масштабируемости и безопасности [11], что отражает динамичное развитие технологии блокчейна и протоколов консенсуса. Таким образом, интеграция методов машинного обучения с блокчейн-технологией представляет собой перспективное направление развития. Современный подход к децентрализации алгоритмов машинного обучения и данных через использование блокчейна 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 LLM с безопасностью и прозрачностью блокчейна, предлагая эффективное синергетическое взаимодействие [12]. Эта интеграция рассматривается через оценку ее преимуществ и вызовов, предлагая архитектурные решения для инкорпорации LLM в децентрализованные приложения, трансформируя неоднозначные намерения в четкие и выполнимые директивы.

Архитектура динамической децентрализованный

Обзор

Веса в предлагаемой архитектуре первоначально устанавливаются с помощью методов обучения с передачей данных, 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 эти локальные модели объединяются в новую, расширенную версию LLM, которая снова распространяется по сети. В основе всего этого процесса лежит блокчейн, облегчающий обновление и контроль версий LLM и включающий криптовалютную схему поощрения для мотивации участия узлов.

Компоненты

Предлагаемая архитектура модели состоит из следующих элементов:

- Nodes: Конечные устройства, которые участвуют в сети, предоставляя вычислительные ресурсы и данные.
- Learning Unit: Обновленные веса модели, полученные узлами в результате локального обучения на своих наборах данных.
- Data Unit: Фрагмент набора данных, который узлы решают передать сети.
- Transaction Unit: Обмен монетами между узлами в сети.
- Memory Pool (Mempool): Место для хранения learning units, data units и transactions перед их компиляцией в новые блоки.
- Transaction Block: Блок, состоящий из нескольких транзакций, выбранных из пула памяти.
- Data Block: Блок, включающий транзакцию coinbase для получения вознаграждения и компиляцию блоков данных из Mempool, формирующих тестовый набор данных.
- Version Block: Блок, содержащий транзакцию coinbase за вознаграждение и агрегированные веса, формирующие глобальную модель, полученную из комбинации единиц обучения из Mempool.
- LinguaChain: Блокчейн, состоящий из transaction blocks, data blocks и version blocks, последовательно связанных между собой с помощью хэшей предыдущих блоков.

Упрощенная схема предлагаемой архитектуры LinguaChain представлена на (см. Рис. 1).

Описание работы

Архитектура начинается с установки начальных весов на выбранном узле путем трансферного обучения, используя веса из предварительно обученной модели, а не случайным выбором. Затем эти веса распространяются по одноранговой сети (P2P), позволяя узлам индивидуально обучать модель на своих данных. После обучения узлы обновляют веса модели и инкапсулируют эти обновления в блок обучения. Этот блок защищается с помощью шифрования с закрытым ключом, создавая уникальную цифровую подпись, и впоследствии отправляется в Mempool. Для поддержки последующего комбинирования моделей применяется гомоморфное шифрование. Одновременно

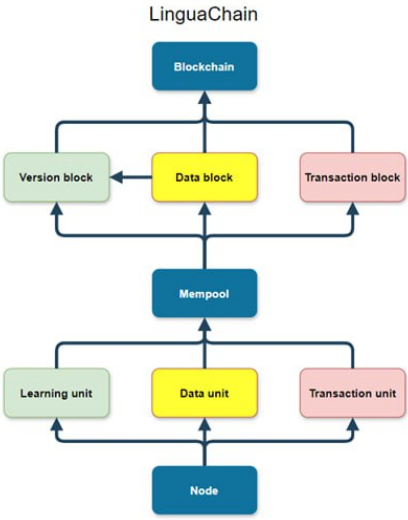

Рис. 1: Архитектура LinguaChain.

узлы совершают криптовалютные транзакции в блокчейне, и эти транзакции также направляются в Mempool. Более того, некоторые узлы предоставляют тестовые наборы данных, шифруя и подписывая единицы данных в процессе, схожем с процессом обучения, перед тем как отправить их в Mempool, где гомоморфное шифрование позволяет выполнять последующие операции. После этого начального этапа механизм Proof-of-Stake (PoS) выбирает узлы на основе их криптовалютных холдингов для создания новых блоков, добавляя определенную степень случайности для обеспечения разнообразного выбора создателей блоков. Эти узлы извлекают, проверяют и собирают транзакции из Mempool в новый блок, который включает транзакцию coinbase [23] в качестве вознаграждения для создателя. Затем этот блок распространяется по сети для проверки целостности. Если большинство узлов подтверждают транзакции, блок интегрируется в блокчейн; 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 и блока версии перед любыми последующими блоками транзакций. Эта процедура использует комбинацию механизмов PoS, Proof-of-Time (PoT) и Proof-of-Work (PoW). Для генерации блока данных PoS определяет выбор узла на основе ставки, при этом назначается случайный вызов (номер), который обрабаты-

вается с помощью функции верифицируемой задержки (VDF). Узлы выбирают и улучшают блоки данных из пула Mempool для повышения качества набора данных. Создатель набора данных наивысшего качества формирует новый блок данных, который затем аутентифицируется сетью с помощью выходных данных VDF. Создатели и авторы проверенных блоков данных получают вознаграждение через транзакцию coinbase, отражающее их вклад в качество набора данных. Узлы, искажающие качество данных или манипулирующие VDF, рискуют лишиться доли. 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. Используя механизм Proof- of-Stake (PoS), узлы выбираются для создания блока версии. 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. Затем эти партии распределяются между выбранными узлами. Выдача вызова вместе с этими разнообразными партиями данных побуждает узлы применять усреднение с помощью Federated Averaging (FedAvg) или аналогичными методами для уточнения агрегированной модели на основе полученных данных. После завершения процесса верификации с помощью Verifiable Delay Function (VDF) узлу, достигшему наивысшей производительности по всему набору тестовых данных, поручается создание блока новой версии. После создания этот блок проходит верификацию во всей сети. Успешный процесс проверки не только подтверждает правильность блока версий, но и вызывает вознаграждение как для создателя блока версий, 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 блоке Version 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 разработана таким образом, чтобы сбалансировать вклад и выгоду 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. Для обеспечения целостности и конфиденциальности данных в этой архитектуре используются безопасные методы агрегирования, такие как гомоморфное шифрование единиц обучения. Также рассматриваются методы дифференциальной конфиденциальности, которые вводят случайный шум в каждую единицу обучения для дальнейшего усиления защиты конфиденциальности.

Заключение

В данной статье описывается теоретическая архитектура, объединяющая LLM и блокчейн для создания масштабируемого и демократизированного ландшафта ИИ, одновременно решая проблемы конфиденциальности и централизованного контроля. Потенциал такого слияния огромен, оно может повысить эффективность LLM и расширить доступ.

Список литературы

[1] X. You, X. Liu, X. Lin, J. Cai, S. Chen, ”Accuracy Degrading: Toward Participation-Fair Federated Learning” in IEEE Internet of Things Journal, June 2023 DOI: 10.1109/JIOT.2023.3238038.

[2] A. Bose, L. Bai, ”A Fully Decentralized Homomorphic Federated Learning Framework” in IEEE 20th International Conference on Mobile Ad Hoc and Smart Systems (MASS), September 2023, DOI: 10.1109/MASS58611.2023.00029.

[3] G. Yu, X. Wang, C. Sun, Q. Wang, P. Yu, W. Ni, R. Liu, X. Xu, ”IronForge: An Open, Secure, Fair, Decentralized Federated Learning”, January 2023, arXiv:2301.04006v1.

[4] Z. Tang, Y. Wang, X. He, L. Zhang, X. Pan, Q. Wang, R. Zeng, K. Zhao, S. Shi, B. He, X. Chu, ”FusionAI: Decentralized Training and Deploying LLMs with Massive Consumer-Level GPUs” in Symposium on Large Language Models (LLM 2023) with IJCAI 2023, Macao, China, August 21, 2023, arXiv:2309.01172.

[5] C. Chen, X. Feng, J. Zhou, J. Yin, X. Zheng, ”Federated Large Language Model: A Position Paper”, Zhejiang University, Hangzhou, China, 18 Jul 2023, arXiv:2307.08925v1.

[6] J. Zhao, Y. Song, S. Liu, I. Harris, S. Jyothi, ”LinguaLinked: A Distributed Large Language Model”, Dec 2023, arXiv:2312.00388v1.

[7] Y. Sheng, L. Zheng, B. Yuan, Z. Li, M. Ryabinin, D. Y. Fu, Z. Xie, B. Chen, C. Barrett, J. Gonzalez, P. Liang, C. Re, I. Stoica, C. Zhang, ”FlexGen: High-Throughput Generative Inference of Large Language Models with a Single GPU”, Jun 2023, arXiv:2303.06865v2.

[8] . Alizadeh, I. Mirzadeh, D. Belenko, S. Khatamifard, M. Cho, C. Mundo, M. Rastegari, M. Farajtabar, ”LLM in a flash: Efficient Large Language Model Inference with Limited Memory”, Jan 2024, arXiv:2312.11514v2.

[9] C. Nguyen, H. Thai, D. Nyato, N. Nguyen, E. Dutkiewicz, ”Proof-of- Stake Consensus Mechanisms for Future Blockchain Networks”, IEEE Access, June 2019, DOI: 10.1109/ACCESS.2019.2925010.

[10] Y. Xu, X. Yang, J. Zhang, J. Zhu, M. Sun, B. Chen, ”Proof of Engagement: A Flexible Blockchain Consensus Mechanism” in Wireless Communications and Mobile Computing, Hindawi, August 2021, DOI: 10.1155/2021/6185910.

[11] T. Do, T. Nguyen, H. Pham, ”Delegated Proof of Reputation: a novel Blockchain consensus”, December 2019, arXiv:1912.04065v1.

[12] X. Liu, ”Decentralized Machine Learning on a Blockchain: Case Studies” in 2023 IEEE International Conference on Omni-layer Intelligent Systems (COINS), July 2023, DOI: 10.1109/COINS57856.2023.10189230.