УДК: 004.89

Худайберидева Г. Б., магистр, ассистент кафедры
«Информатика и информационные технологии»
Московский Политехнический Университет,
Россия, г. Москва

Кожухов Д. А., магистр, ассистент кафедры
«Информатика и информационные технологии»
Московский Политехнический Университет,
Россия, г. Москва

Пименкова А. А., студент-бакалавр кафедры
«Информатика и информационные технологии»
Московский Политехнический Университет,
Россия, г. Москва

ДИНАМИЧЕСКОЕ УПРАВЛЕНИЕ ТОЧНОСТЬЮ И СЛОЖНОСТЬЮ В РЕАЛЬНОМ ВРЕМЕНИ НА ОСНОВЕ ДОСТУПНОЙ МОЩНОСТИ В МИКРО-LLM

***Аннотация**: Рассматривается проблема энергопотребления больших языковых моделей (LLM) при развертывании на устройствах с батарейным питанием и строгими энергетическими ограничениями. Предлагается концепция микро-LLM, оснащенных механизмами динамической адаптации своей вычислительной сложности и числовой точности в реальном времени, основанной на текущем уровне доступной мощности или заданном пользователем энергетическом бюджете. Ключевыми аспектами инновации являются методы селективной активации компонентов модели (слоев, голов внимания), адаптации битовой ширины вычислений и специализированного рантаим-менеджмента для управления энергопотреблением. Анализируются требования к архитектуре модели, рантайм-системе и потенциальные выгоды в контексте энергоэффективности. Указываются основные технические вызовы, требующие решения для практической реализации.*

Ключевые слова: *большие языковые модели, микро-LLM, энергоэффективность, динамическая адаптация, управление мощностью, квантование, аппаратно-программная кооперация, ресурсоограниченные устройства, рантайм-менеджмент, батарейное питание.*

Khudaiberideva G. B.
master and department assistant at the department of
"Computer Science and Information Technology"
Moscow Polytechnic University
Moscow, Russia

Kozhukhov D. A.
master and department assistant at the department of
"Computer Science and Information Technology"
Moscow Polytechnic University
Moscow, Russia

Pimenkova A. A.
bachelor's student at the department of
"Computer Science and Information Technology"
Moscow Polytechnic University
Moscow, Russia

DYNAMIC CONTROL OF ACCURACY AND COMPLEXITY IN REAL TIME BASED ON AVAILABLE POWER IN MICRO-LLM

Annotation:*The problem of energy consumption of large language models (LLM) when deployed on battery-powered devices with strict energy constraints is considered. The concept of micro-LLMs is proposed, equipped with mechanisms for dynamically adapting their computational complexity and numerical accuracy in real time, based on the current level of available power or a user-defined energy budget. The key aspects of innovation are methods of selective activation of model components (layers, heads of attention), adaptation*

of the bit width of calculations and specialized rankaim management for energy consumption management. The requirements for the architecture of the model, the runtime system and the potential benefits in the context of energy efficiency are analyzed. The main technical challenges that require solutions for practical implementation are indicated.

***Keywords**: large language models, micro-LLM, energy efficiency, dynamic adaptation, power management, quantization, hardware and software cooperation, resource-limited devices, runtime management, battery power.*

Введение

Широкое внедрение больших языковых моделей сталкивается с существенным барьером в виде их высоких требований к вычислительным ресурсам и энергопотреблению [1, 4]. Традиционные LLM, демонстрирующие высокую производительность, часто неприменимы на устройствах с батарейным питанием (мобильные устройства, носимые гаджеты, IoT-платформы) или в системах с жесткими энергетическими лимитами [11]. Энергопотребление становится критическим фактором, определяющим возможность развертывания и продолжительность автономной работы [4, 9]. Статическая оптимизация моделей для таких сред, хотя и является важным направлением [5, 6], не учитывает изменчивость доступной мощности в реальных условиях эксплуатации. Источник энергии (сеть, батарея с изменяющимся уровнем заряда, суперконденсатор) и приоритеты пользователя (максимальная производительность vs продление времени работы) создают динамический контекст, требующий адаптивного подхода [13]. Возникает потребность в принципиально новых методах управления ресурсами LLM в реальном времени [8, 12].

Постановка Проблемы Энергопотребления LLM

Энергопотребление LLM напрямую коррелирует с объемом выполняемых вычислений, определяемым размером модели и битовой шириной операций [4, 10]. Трансформаторная архитектура, лежащая в основе современных LLM [1], характеризуется значительными затратами

энергии на операции матричного умножения и внимания [9, 10]. Уменьшение размера модели (создание микро-LLM) [5, 6] и применение квантования (8-bit, 4-bit) [3, 14, 15] являются стандартными методами снижения энергозатрат. Однако эти подходы носят статический характер. Модель, оптимизированная для работы от батареи, не сможет использовать избыточную мощность от сети для повышения точности. И наоборот, модель, работающая с высокой точностью при питании от сети, может превысить допустимый энергобюджет при переходе на батарею, приводя к неконтролируемому завершению работы [11, 13]. Отсутствие механизмов динамического масштабирования вычислительной нагрузки модели в ответ на изменения доступной мощности представляет собой существенный пробел [8, 12].

Концепция Динамической Адаптации Микро-LLM

Предлагаемая инновация заключается в разработке микро-LLM со встроенной способностью к динамическому изменению своей вычислительной сложности и числовой точности во время исполнения (inference) [5, 13]. Изменение сложности подразумевает адаптацию глубины модели, выражаемую в количестве активных трансформаторных слоев [2]. Альтернативно или дополнительно может адаптироваться ширина модели через активацию или деактивацию части голов механизма внимания в слоях [2, 17]. Изменение точности достигается переключением между разными режимами квантования весов и активаций модели (например, между 8-bit и 4-bit представлениями) в процессе работы [3, 7, 14, 15]. Ключевым принципом является прямая зависимость выбора режима работы от текущего уровня доступной мощности, измеряемого системой мониторинга [8, 9], или от явно заданного пользователем энергетического бюджета [13]. Целью является максимизация полезного выхода модели (например, качества генерируемого текста) в рамках жесткого и динамически меняющегося энергетического ограничения [4, 8].

Архитектурные Аспекты и Рантайм-Менеджмент

Реализация концепции требует глубокой интеграции на уровне архитектуры модели, аппаратного обеспечения и специализированного программного рантайм-менеджмента [10, 12, 16]. Микро-LLM должна быть спроектирована с поддержкой модульности и возможности изолированного отключения компонентов [2, 6, 17]. Это предполагает введение механизмов "байпаса" для слоев и возможность условного выполнения групп вычислений [2, 16]. Аппаратная платформа должна обеспечивать эффективное измерение потребляемой мощности в реальном времени [9] и предоставлять интерфейсы для быстрого изменения режимов работы вычислительных блоков, включая переключение между целочисленными блоками разной битности [15, 16]. Рантайм-менеджер выступает центральным звеном системы [8, 12, 13]. Его функции включают непрерывный мониторинг доступной мощности (посредством датчиков или системных интерфейсов) [9] и текущего энергопотребления модели, прогнозирование затрат для различных конфигураций сложности/точности [8], принятие решений о переключении режима на основе заданной политики (например, максимизация качества при заданном бюджете или минимизация потребления при заданном минимальном качестве) [8, 13] и выполнение самого переключения с минимальными накладными расходами [12, 16]. Рантайм-менеджер должен обладать знанием энергетического профиля каждого возможного состояния модели [8, 9].

Технические Вызовы

Разработка динамически адаптивных микро-LLM сопряжена с рядом значительных технических вызовов. Проблема эффективного и быстрого переключения между состояниями модели требует решения [12, 16]. Переход между разными уровнями квантования может потребовать перезагрузки весов в оперативную память или переконфигурации

вычислительных ядер [3, 15]. Активация и деактивация слоев или голов внимания должна происходить без нарушения целостности состояния модели [2, 17]. Обеспечение плавности переключения и минимизация задержки являются критическими задачами [12, 16]. Создание точных и легковесных моделей энергопотребления для различных конфигураций микро-LLM представляет отдельную сложность [8, 9]. Эти модели должны учитывать не только объем вычислений, но и энергозатраты на доступ к памяти и коммуникацию [4, 9, 10]. Разработка алгоритмов принятия решений рантайм-менеджером, оптимально балансирующих качество вывода и энергопотребление в условиях неопределенности и динамики, требует применения продвинутых методов оптимизации и, возможно, машинного обучения [8, 13]. Валидация всего подхода требует создания специализированных бенчмарков, отражающих сценарии с переменным энергобюджетом [4, 11].

Заключение

Динамическое управление точностью и сложностью микро-LLM на основе доступной мощности представляет собой перспективное направление для преодоления ограничений, накладываемых энергопотреблением при развертывании LLM на ресурсоограниченных устройствах. Предложенная концепция предполагает создание моделей и инфраструктуры исполнения, способных адаптировать свои вычислительные требования в реальном времени, реагируя на изменения в доступной энергии или пользовательских предпочтениях относительно энергобюджета. Основой для реализации являются селективная активация компонентов модели , адаптация битовой ширины вычислений и интеллектуальный рантайм-менеджмент. Преодоление связанных технических вызовов, таких как эффективное переключение состояний, точное моделирование энергопотребления и разработка оптимальных политик управления, является необходимым условием для практического

воплощения этой инновации. Успешная реализация позволит существенно расширить область применения LLM, включив в нее широкий спектр портативных и автономных устройств с батарейным питанием.

СПИСОК ЛИТЕРАТУРЫ:

1. Brown, T. et al. Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems 33 (NeurIPS 2020). 2020.
2. Fedus, W., Zoph, B., Shazeer, N. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. Journal of Machine Learning Research (JMLR). 2022. Vol. 23.
3. Dettmers, T. et al. QLoRA: Efficient Finetuning of Quantized LLMs. arXiv preprint arXiv:2305.14314. 2023.
4. Kim, S. et al. Energy-Efficient Transformer Training and Inference for Edge Devices: A Review. IEEE Access. 2023. Vol. 11.
5. Xu, Z. et al. TinyLlama: An Open-Source Small Language Model. arXiv preprint arXiv:2401.02385. 2024.
6. Wu, C. et al. MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL). 2020.
7. Li, Y. et al. Ternary Neural Networks for Resource-Efficient AI Applications. IEEE Transactions on Neural Networks and Learning Systems (TNNLS). 2020. Vol. 31, No. 9.
8. Riera, M. et al. Runtime Energy Estimation and Optimization for Deep Neural Networks. ACM Transactions on Embedded Computing Systems (TECS). 2022. Vol. 21, No. 4.
9. Horowitz, M. Energy Table for 45nm Process. Stanford VLSI Wiki. [Электронный ресурс]. URL: https://vlsiweb.stanford.edu/

10. Sze, V., Chen, Y., Yang, T., Emer, J. Efficient Processing of Deep Neural Networks: A Tutorial and Survey. Proceedings of the IEEE. 2017. Vol. 105, No. 12.
11. Parashar, A. et al. Energy-Efficient Inference Serving for Deep Learning Applications. Proceedings of the 14th ACM International Conference on Future Energy Systems (e-Energy). 2023.
12. Chen, T. et al. Dynamic Voltage and Frequency Scaling for Energy-Efficient GPU Computing. IEEE Transactions on Parallel and Distributed Systems (TPDS). 2023. Vol. 34, No. 4.
13. Benini, L., De Micheli, G. Dynamic Power Management: Design Techniques and CAD Tools. Kluwer Academic Publishers. 1998.
14. Krishnamoorthi, R. Quantizing deep convolutional networks for efficient inference: A whitepaper. arXiv preprint arXiv:1806.08342. 2018.
15. Jacob, B. et al. Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018.
16. Han, S. et al. ESE: Efficient Speech Recognition Engine for Mobile Devices. Proceedings of the 22nd International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS). 2017.
17. Lebedev, V., Lempitsky, V. Fast ConvNets Using Group-wise Brain Damage. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016.