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

АНАЛОГОВЫЕ И ИМПУЛЬСНЫЕ НЕЙРОННЫЕ СЕТИ КАК ОСНОВА ДЛЯ СЛЕДУЮЩЕГО ПОКОЛЕНИЯ СВЕРХ-ЭФФЕКТИВНЫХ МИКРО-LLM В ПРОМЫШЛЕННОСТИ

***Аннотация**: Исследуется потенциал парадигмального сдвига в архитектуре микро-LLM (маломасштабных больших языковых моделей) для промышленных приложений через полный отказ от цифровых вычислений фон Неймана. Анализируется возможность использования аналоговых схем и импульсных нейронных сетей (Spiking Neural Networks, SNN) в качестве фундамента для создания сверх-эффективных решений. Основное внимание уделяется способности данных подходов обеспечить качественное улучшение показателей энергоэффективности и скорости обработки при выполнении простых языковых задач, таких как классификация интентов и извлечение ключевых фраз, в условиях работы на ресурсоограниченных промышленных контроллерах, включая сверх-медленные или аналоговые платформы. Рассматриваются фундаментальные принципы работы SNN и аналоговых систем, их соответствие требованиям промышленной среды, современное состояние*

прототипирования и ключевые исследовательские направления. Делается вывод о существенном, хотя и сопряженном с техническими сложностями, потенциале данных технологий для создания нового класса микро-LLM.

***Ключевые слова**: импульсные нейронные сети (SNN), аналоговые вычисления, микро-LLM, энергоэффективность, промышленные контроллеры, архитектура фон Неймана, классификация интентов, извлечение ключевых фраз, нейроморфные системы.*

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

ANALOG AND PULSE NEURAL NETWORKS AS THE BASIS FOR THE NEXT GENERATION OF ULTRA-EFFICIENT MICRO-LLMS IN INDUSTRY

***Annotation**:The potential of a paradigm shift in the architecture of micro-LLM (small-scale large language models) for industrial applications through the complete*

abandonment of von Neumann digital computing is being investigated. The possibility of using analog circuits and pulsed neural networks (Spiking Neural Networks, SNN) as a foundation for creating ultra-efficient solutions is analyzed. The main focus is on the ability of these approaches to provide qualitative improvements in energy efficiency and processing speed when performing simple language tasks such as classifying intents and extracting keywords, when operating on resource-limited industrial controllers, including ultra-slow or analog platforms. The fundamental principles of operation of SNN and analog systems, their compliance with the requirements of the industrial environment, the current state of prototyping and key research areas are considered. It is concluded that the potential of these technologies for creating a new class of micro-LLMs is significant, although fraught with technical difficulties.

***Keywords**: pulsed neural networks (SNN), analog computing, micro-LLM, energy efficiency, industrial controllers, von Neumann architecture, intent classification, keyword extraction, neuromorphic systems.*

Введение

Распространение технологий искусственного интеллекта в промышленной автоматизации сталкивается с фундаментальным ограничением, связанным с архитектурой фон Неймана [1], доминирующей в современных цифровых вычислительных системах. Разделение процессора и памяти создает узкое место, известное как "бутылочное горлышко фон Неймана", приводящее к значительным затратам энергии на перемещение данных, особенно при работе с моделями, требующими интенсивных вычислений, даже в их уменьшенных формах (микро-LLM) [2]. Промышленные контроллеры, особенно устаревшие или специализированные сверх-медленные либо аналоговые системы, обладают крайне ограниченными вычислительными ресурсами и энергетическим бюджетом [3]. Традиционные микро-LLM, основанные на цифровых нейронных сетях, часто оказываются неприменимы на таких платформах из-за высоких требований к производительности и энергии. Возникает потребность в принципиально новых подходах к архитектуре вычислений для микро-LLM, способных

функционировать в жестких промышленных условиях. Аналоговые схемы и импульсные нейронные сети представляют собой альтернативные парадигмы, черпающие вдохновение в биологических нейронных системах и предлагающие путь к преодолению ограничений фон Неймановской архитектуры [4].

Ограничения Архитектуры фон Неймана для Микро-LLM в Промышленности

Эффективное развертывание микро-LLM на промышленных контроллерах наталкивается на непреодолимые барьеры, порожденные самой природой цифровых вычислений. Цифровые процессоры выполняют операции последовательно над дискретными значениями, представленными бинарным кодом. Каждая операция требует извлечения инструкций и данных из памяти, их декодирования, выполнения и последующего сохранения результата обратно в память [1]. Этот цикл неизбежно связан с перемещением больших объемов данных по шинам памяти, что является энергетически наиболее затратной частью вычислений в современных чипах [5]. Микро-LLM, даже оптимизированные, требуют выполнения миллионов операций умножения-накопления (MAC) для обработки даже простых языковых конструкций, таких как определение намерения пользователя в команде управления или выделение ключевых параметров из технического сообщения [6]. На сверх-медленных контроллерах с тактовой частотой в единицы или десятки мегагерц выполнение таких задач цифровым способом приводит к неприемлемым задержкам. Аналоговые контроллеры, изначально не предназначенные для сложной цифровой обработки, лишены необходимой вычислительной базы для запуска традиционных нейросетевых алгоритмов. Энергетический бюджет многих промышленных устройств, особенно работающих от автономных источников питания или в удаленных местах, измеряется микроваттами или милливаттами, что

исключает использование энергоемких цифровых сопроцессоров [3]. Следовательно, использование микро-LLM на основе фон Неймановской архитектуры в рассматриваемых условиях промышленного применения либо невозможно, либо крайне неэффективно.

Импульсные Нейронные Сети (SNN): Биологическое Вдохновение и Принципы

Импульсные нейронные сети предлагают радикально иную модель вычислений, основанную на принципах работы биологического мозга [7]. В отличие от искусственных нейронов в традиционных глубоких сетях, которые передают непрерывные значения активации на каждом временном шаге, нейроны в SNN обмениваются дискретными электрическими импульсами, или спайками, в асинхронные моменты времени [8]. Информация в SNN кодируется не амплитудой сигнала, а временными паттернами спайков (временным кодом), частотой следования спайков (частотным кодом) или их латентностью [9]. Состояние нейрона моделируется его мембранным потенциалом. Приходящие спайки от пресинаптических нейронов изменяют этот потенциал. Когда мембранный потенциал достигает определенного порога, нейрон генерирует собственный спайк и сбрасывает свой потенциал [10]. Ключевым аспектом SNN является использование динамики во времени. Обработка информации происходит асинхронно и событийно: вычисления активируются только при поступлении спайков, а не на каждом такте синхронизации, как в цифровых системах [11]. Обучение SNN часто опирается на правила, учитывающие временные зависимости между спайками, такие как Spike-Timing-Dependent Plasticity (STDP), которые моделируют синаптическую пластичность в биологических системах [12]. Данные свойства принципиально отличают SNN от традиционных искусственных нейронных сетей и создают предпосылки для высокой

энергоэффективности при реализации в специализированном аппаратном обеспечении.

Аналоговые Вычисления: Непрерывность и Параллелизм

Аналоговые вычисления оперируют непрерывными физическими величинами, такими как напряжение или ток, для непосредственного представления и обработки данных [13]. Математические операции, фундаментальные для нейронных сетей, такие как умножение и сложение, могут выполняться аналоговыми компонентами (например, операционными усилителями, транзисторами в определенных режимах) с высокой скоростью и крайне низким энергопотреблением [14]. В аналоговой нейроморфной системе синаптические веса могут быть представлены проводимостями мемристоров или другими аналоговыми элементами памяти, а активации нейронов – уровнями напряжения или тока [15]. Важнейшим преимуществом аналогового подхода является возможность реализации операций умножения-накопления in-memory или in-situ [16]. Это означает, что вычисления происходят непосредственно в месте хранения весовых коэффициентов, кардинально снижая или полностью устраняя необходимость перемещения данных между памятью и процессором, что является главным источником энергозатрат в фон Неймановских системах [5]. Аналоговые системы обладают высокой степенью параллелизма, так как множество операций может выполняться одновременно через физические законы, управляющие электрическими цепями [17]. Данные характеристики делают аналоговые схемы потенциально идеальными кандидатами для реализации энергоэффективных микро-LLM.

Потенциал SNN и Аналоговых Схем для Микро-LLM в Промышленности

Синергия принципов SNN и аналоговой реализации открывает путь к созданию микро-LLM с беспрецедентной энергоэффективностью и скоростью, критически важными для промышленных применений. Событийно-управляемая природа SNN означает, что энергия потребляется

только при обработке входящего спайка, а не постоянно, как в синхронных цифровых схемах [11]. При реализации на аналоговом или смешанном аналого-цифровом оборудовании, операции обработки спайков могут выполняться с минимальными затратами энергии за счет использования физических свойств электронных компонентов [14, 17]. Теоретически, такие системы могут приближаться к энергоэффективности биологического мозга [4]. Высокая скорость аналоговых вычислений позволяет обрабатывать временные паттерны спайков с задержками, существенно меньшими, чем время, необходимое для выполнения эквивалентных цифровых операций на медленных контроллерах [13]. Что касается применимости к языковым задачам, исследования демонстрируют возможность эффективного использования SNN для классификации текстовых последовательностей и извлечения паттернов [9]. Преобразование текста во временные или частотные коды спайков является решающим шагом. Простые языковые задачи, такие как распознавание ограниченного набора интентов (например, "запустить", "остановить", "статус", "авария") или извлечение ключевых фраз (например, именованных сущностей: названия агрегатов, коды ошибок, числовые параметры), часто опираются на распознавание шаблонов и могут быть эффективно смоделированы с помощью обученных SNN соответствующей архитектуры [6]. Способность SNN и аналоговых систем работать с шумом и в условиях неидеальности компонентов [15] также коррелирует с требованиями промышленной среды. Интеграция таких микро-LLM могла бы осуществляться через специализированные сопроцессоры, взаимодействующие с основным промышленным контроллером по низкоскоростным промышленным шинам (например, Modbus, CAN), или путем полной аналоговой реализации логики обработки команд.

Фундаментальные Исследования и Практические Прототипы

Исследовательская деятельность в области применения SNN и аналоговых вычислений для задач, родственных микро-LLM, активно развивается, хотя и фокусируется преимущественно на более простых задачах, чем полная генерация текста. Фундаментальные исследования охватывают разработку эффективных алгоритмов преобразования текстовых данных в спайковые последовательности (кодирование) [9], создание новых и более обучаемых моделей аналоговых и спайковых нейронов [10, 12], а также исследование архитектур сетей, оптимальных для обработки последовательностей символов или семантических признаков в спайковом представлении [8]. Значительные усилия направлены на разработку алгоритмов обучения SNN, которые были бы эффективны и реализуемы на аппаратном уровне, включая варианты обратного распространения ошибки, адаптированные для временных кодов, и методы обучения без учителя на основе STDP [12]. В области практического прототипирования наблюдаются несколько направлений. Разрабатываются специализированные нейроморфные чипы, такие как Loihi от Intel (цифровой, но имитирующий SNN), BrainScaleS (аналоговый) [4], а также чипы компаний Mythic AI и Rain Neuromorphics, использующие аналоговые in-memory вычисления [16, 17]. Эти платформы демонстрируют на порядки более высокую энергоэффективность при выполнении задач классификации паттернов и обработки сенсорных данных по сравнению с традиционными GPU/CPU [5]. Хотя большинство демонстраций сосредоточено на обработке изображений или аудио, принципы обработки временных последовательностей применимы и к текстовым данным. Прототипы систем для простой классификации текста или обработки событий на основе SNN существуют в академической среде [8, 9], подтверждая принципиальную осуществимость подхода для задач, актуальных для микро-LLM. Разработка специализированных аналоговых или SNN-сопроцессоров, интегрируемых в промышленные шины данных,

представляется логичным следующим шагом для прикладных исследований.

Технические Вызовы и Направления Исследований

Несмотря на значительный потенциал, переход к микро-LLM на основе SNN и аналоговых схем сопряжен с серьезными техническими вызовами. Точность аналоговых компонентов подвержена влиянию шумов, температурных дрейфов и вариаций технологического процесса изготовления [15]. Обеспечение достаточной точности и воспроизводимости вычислений для языковых задач, где даже небольшие ошибки могут исказить смысл, требует разработки устойчивых архитектур и схем компенсации. Эффективное кодирование текстовой информации в пространственно-временные паттерны спайков для задач сложнее простой классификации остается открытой проблемой [9]. Обучение крупных SNN, особенно с использованием временных кодов, является вычислительно сложной задачей, часто требующей ресурсоемкой симуляции или специализированного оборудования [12]. Перенос обученных моделей на физические аналоговые или SNN-платформы с их аппаратными неидеальностями требует разработки методик калибровки и адаптации. Интеграция нейроморфных или аналоговых сопроцессоров со стандартными промышленными контроллерами и шинами данных требует решения вопросов интерфейсов, синхронизации и управления. Основные направления фундаментальных и прикладных исследований включают разработку устойчивых аналоговых элементов памяти и нейронных схем [14, 15], создание эффективных и аппаратно-реализуемых алгоритмов обучения SNN для последовательностей [12], исследование новых методов кодирования семантической информации в спайковые потоки [9], проектирование специализированных архитектур SNN для конкретных промышленных языковых задач [6], а также разработку методологий

совместного проектирования аппаратного и программного обеспечения (hardware-software co-design) для промышленных микро-LLM [17].

Заключение

Анализ фундаментальных принципов работы импульсных нейронных сетей и аналоговых вычислений, а также современного состояния исследований и прототипирования позволяет сделать вывод о значительном потенциале этих технологий для создания следующего поколения микро-LLM, ориентированных на промышленные применения. Отказ от парадигмы фон Неймана в пользу событийно-управляемых, асинхронных вычислений на основе спайков и непрерывных физических величин предлагает путь к преодолению ключевых ограничений по энергопотреблению и скорости, делающих традиционные микро-LLM неприменимыми на сверх-медленных или аналоговых промышленных контроллерах. Принципиальная возможность эффективного выполнения SNN простых, но важных языковых задач, таких как классификация интентов и извлечение ключевых фраз, подтверждается существующими исследованиями в области обработки временных последовательностей на нейроморфных платформах. Демонстрации на прототипах аналоговых и нейроморфных чипов показывают достижимость на порядки более высокой энергоэффективности по сравнению с цифровыми системами. Однако реализация этого потенциала в промышленных микро-LLM требует преодоления существенных технических вызовов, связанных с точностью аналоговых компонентов, сложностью кодирования и обучения SNN для языковых задач, а также вопросами интеграции. Последующие фундаментальные исследования должны быть сосредоточены на разработке устойчивых аналоговых схем, эффективных алгоритмов обучения SNN для последовательностей и методов семантического кодирования.

СПИСОК ЛИТЕРАТУРЫ:

1. Ambrogio S. et al. Equivalent-accuracy accelerated neural-network training using analogue memory // Nature. 2018. Vol. 558, № 7708. P. 60–67.
2. Chua L. Memristor-The missing circuit element // IEEE Transactions on circuit theory. 1971. Vol. 18, № 5. P. 507–519.
3. Diehl P.U., Cook M. Unsupervised learning of digit recognition using spike-timing-dependent plasticity // Frontiers in Computational Neuroscience. 2015. Vol. 9. P. 99.
4. Gungor V.C., Hancke G.P. Industrial Wireless Sensor Networks: Challenges, Design Principles, and Technical Approaches // IEEE Transactions on Industrial Electronics. 2009. Vol. 56, № 10. P. 4258–4265.
5. Horowitz M. 1.1 Computing's energy problem (and what we can do about it) // 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC). IEEE, 2014. P. 10–14.
6. Izhikevich E.M. Simple model of spiking neurons // IEEE Transactions on neural networks. 2003. Vol. 14, № 6. P. 1569–1572.
7. Li C. et al. Analogue signal and image processing with large memristor crossbars // Nature Electronics. 2018. Vol. 1, № 1. P. 52–59.
8. Liu X. et al. Tiny Machine Learning: Progress and Futures // IEEE Circuits and Systems Magazine. 2022. Vol. 22, № 3. P. 8–33. DOI: 10.1109/MCAS.2022.3197502.
9. Maass W. Networks of spiking neurons: the third generation of neural network models // Neural Networks. 1997. Vol. 10, № 9. P. 1659–1671.
10. Mead C. Analog VLSI and neural systems. Reading, Mass.: Addison-Wesley, 1989. 371 p.

11. Ponulak F., Kasiński A. Introduction to spiking neural networks: Information processing, learning and applications // Acta neurobiologiae experimentalis. 2011. Vol. 71, № 4. P. 409–433.
12. Roy K., Jaiswal A., Panda P. Towards spike-based machine intelligence with neuromorphic computing // Nature. 2019. Vol. 575, № 7784. P. 607–617.
13. Schuman C.D. et al. A Survey of Neuromorphic Computing and Neural Networks in Hardware // arXiv preprint arXiv:1705.06963. 2022.
14. Tavanaei A. et al. Deep learning in spiking neural networks // Neural Networks. 2019. Vol. 111. P. 47–63.
15. Von Neumann J. First Draft of a Report on the EDVAC. 1945.
16. Wu C.J. et al. Sustainable AI: Environmental Implications, Challenges and Opportunities // Proceedings of Machine Learning and Systems. 2022. Vol. 4. P. 795–813.
17. Yao P. et al. Fully hardware-implemented memristor convolutional neural network // Nature. 2020. Vol. 577, № 7792. P. 641–646.