

Аппаратное исполнение импульсных нейронных сетей в нейроморфных электронных модулях

А. П. Дудкин, Н. В. Андреева, Е. А. Рындин

Санкт-Петербургский государственный электротехнический университет
«ЛЭТИ» им. В.И. Ульянова (Ленина)

anddudkin000@gmail.com, nvandr@gmail.com, rynenator@gmail.com

Аннотация. Разработаны модель, методика и программные средства моделирования импульсных нейронных сетей, позволяющие учесть и оценить влияние особенностей мемристивной электронной компонентной базы и архитектуры кроссбар-массива при аппаратном исполнении нейронной сети на точность их работы. Получены результаты тестирования импульсной нейронной сети в режиме обучения и инференс-режиме в задаче распознавания изображений с учетом характеристик экспериментально изготовленных мемристивных структур.

Ключевые слова: импульсные нейронные сети, аппаратное исполнение, мемристивный кроссбар-массив

I. ВВЕДЕНИЕ

За последние несколько лет развитие глубоких нейронных сетей и спектр решаемых ими задач достигли впечатляющих результатов. Однако, с увеличением сложности моделей нейросетей, значительно вырастают требования к производительности и энергоэффективности вычислительных устройств [1], что приводит к необходимости поиска методов оптимизации алгоритмов работы нейросетей или разработке альтернативных подходов для более эффективного решения задач машинного обучения.

Импульсные нейронные сети (ИмНС) – это новое поколение нейросетей, обработка информации в которых основана на биологически реалистичных механизмах функционирования головного мозга (рис. 1а) – наиболее «эффективного вычислительного устройства» с точки зрения энергопотребления и производительности [2].

Исполнение ИмНС нецелесообразно на базе вычислительных устройств с архитектурой фон Неймана, проблема которой заключается в пространственном разделении блоков памяти и вычислительных блоков. Концепция вычислений в памяти (in-memory computing, CIM) является перспективным архитектурным решением, позволяющим реализовать преимущество ИмНС. В CIM-архитектурах вычисления и хранение информации осуществляется в синапсах. Максимальной эффективности при разработке нейроморфных электронных модулей можно достичь за счет использования в качестве синапсов новой мемристивной электронной компонентной базы (ЭКБ), аппаратно интегрируемой в структуре кроссбар-массива (рис. 1б).

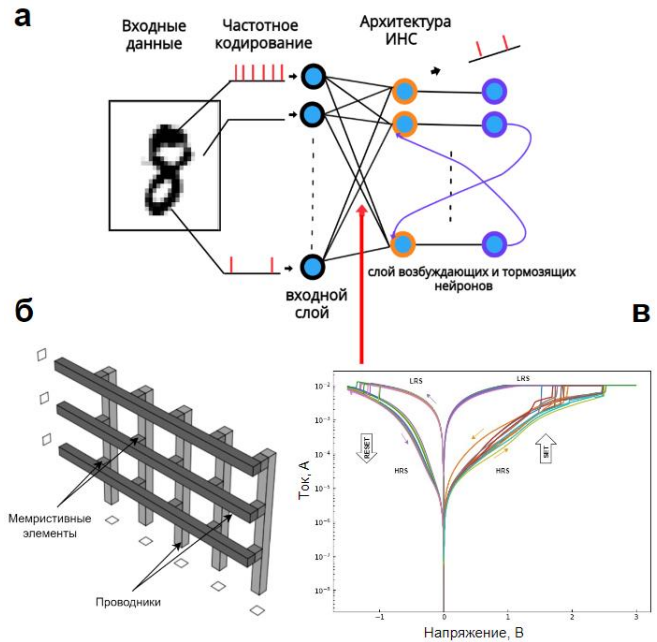


Рис. 1. Принцип работы ИмНС (а), структура мемристивного кроссбар-массива (б), экспериментальные ВАХ мемристора на основе наноразмерной пленки TiOx (в)

Мемристивный кроссбар-массив — это структура из ортогональных проводящих линий металлизации, на пересечении которых находятся мемристивные элементы (композиция наноразмерных оксидных пленок [3]), каждый из которых выступает в роли элемента энергонезависимой памяти (рис. 1б) и выполняет функцию синапса. На строки кроссбара подается вектор входных напряжений V , а со столбцов считывается вектор выходных токов I , являющийся результатом матрично-векторного умножения (МВУ), вычисляемым за один проход в аналоговом виде по законам Ома и Кирхгофа. Подобный подход значительно ускоряет вычисления МВУ в сравнении с программными методами расчета.

Описанный выше принцип функционирования мемристивного кроссбар-массива верен в первом приближении, однако в реальных кроссбар-массивах необходимо учитывать ряд факторов, которые будут влиять на результат МВУ:

- соединительные линии имеют определенное сопротивление, и часть приложенных напряжений будет падать на отрезках соединений;

- реальные мемристивные структуры имеют дискретный диапазон возможных состояний проводимости;
- ВАХ мемристивных элементов нелинейны (рис. 1б).

Все перечисленные факторы должны учитываться в процессе проектирования, так как будут влиять на результат МВУ, и соответственно на результат работы ИмНС при переходе от программной модели к аппаратному исполнению. В работах [4–6] проведена программная симуляция некоторых особенностей аппаратного исполнения ИмНС с использованием мемристивных кроссбар-массивов, и результаты показывают, что эффективность работы нейронных сетей при практической реализации может снижаться, в худшем случае, более чем на 20 % в сравнении с идеализированным программным исполнением.

Целью данной работы является исследование возможностей перехода от абстрактной программной модели ИмНС к аппаратному исполнению нейросети с интеграцией мемристивной ЭКБ и оценка влияния реальных физических характеристик мемристивных кроссбар-массивов на режим обучения и инференс-режим работы ИмНС.

II. МОДЕЛЬ ИМПУЛЬСНОЙ НЕЙРОННОЙ СЕТИ

A. Представление информации

Принимая во внимание тот факт, что ИмНС работают посредством передачи импульсов, информация, передаваемая в нейросеть, должна быть в соответствующем импульсном формате.

В рамках проведенного исследования для обучения и тестирования ИмНС использовалась база данных изображений рукописных цифр MNIST [7]. Изображения имеют размерность 28 на 28 пикселей в оттенках серого (каждому пикселю соответствует значения от 0 до 255). Для ускорения вычислений размерность изображений была уменьшена до 14 на 14 пикселей с помощью метода интерполяции по ближайшему соседу [8].

Для конвертации изображений в импульсный формат использован метод частотного кодирования.

Каждому входу ИмНС сопоставляется один из пикселей изображения, таким образом, вероятность генерации импульса нейроном входного слоя определяется значением интенсивности пикселя изображения.

B. Архитектура ИмНС

В экспериментах использовалась ИмНС с архитектурой двухслойного перцептрона (рис. 1а) со входным слоем из 196 импульсных нейронов (соответствует размерности изображений 14x14 пикселей) и выходным слоем из 50 интегрирующих нейронов с утечкой и адаптивным порогом срабатывания. Нейроны различных слоев соединены между собой синапсами по принципу «каждый с каждым».

Входной слой нейронов не обладает какими-либо специализированными параметрами и выполняет функцию передачи входящей импульсной последовательности на следующий слой нейронов.

Нейроны выходного слоя обрабатывают приходящие по синапсам импульсы и накапливают мембранный потенциал, который при достижении порогового значения заставляет нейрон сгенерировать импульс и уйти в неактивное состояние на определённый промежуток времени – рефрактерный период.

Динамика изменения мембранного потенциала и адаптивного порога срабатывания описываются следующими уравнениями:

$$U_{mem} = (U_{mem} + \sum \frac{I_n}{G_n}) * \tau \text{ для } U_{mem} < U_{trn}$$

$$U_{mem} = U_{rest} \text{ для } U_{mem} \geq U_{trn}$$

$$U_{trn} = U_{trn} + k \text{ для } U_{trnMIN} < U_{trn} < U_{trnMAX}$$

$$U_{trn} = U_{trn} * \tau_{tr} \text{ для } U_{trn} > U_{trnMIN}$$

где, U_{mem} – мембранный потенциал нейрона; I_n – сигнал, приходящий с n -ого нейрона; G_n – проводимости n -ого синапса; U_{trn} – порог срабатывания для n нейрона; k – константа, на которую увеличивается порог срабатывания нейрона после генерации импульса; τ_{tr} – постоянная релаксации порога срабатывания нейронов ($\tau_{tr} < 1$); U_{rest} – потенциал покоя, в котором находится нейрон после активации (рефрактерный период); U_{trnMIN} , U_{trnMAX} – минимальное и максимальное значение порога срабатывания; τ – постоянная релаксации мембранного потенциала.

Адаптивный порог срабатывания необходим для регулирования активности нейронов в процессе обучения ИмНС. Каждая генерация выходного импульса нейрона снижает его активность за счет увеличения порога срабатывания.

C. Методика обучения

Для обучения ИмНС в основном используются локальные правила обучения, что лучше согласуется с известными биологическими механизмами и дискретной природой событий, происходящих в ИмНС. В рамках данной работы для обучения использовалось локальное правило обучения, основанное на синаптической пластичности нейронов (STDP, spike-timing-dependent plasticity).

Обучение происходит за счет усиления или ослабления весов синапсов, которые зависят от времени прихода постсинаптических и пресинаптических импульсов нейрону и описываются следующими зависимостями:

$$\Delta w = A_+ * e^{\frac{t_{pre} - t_{post}}{\tau}} \text{ для } t_{post} > t_{pre} \text{ и } \Delta t < t_+$$

$$\Delta w = A_- \text{ для } t_{post} < t_{pre} \text{ и } \Delta t < t_-$$

где, t_{pre} – время прихода импульса от пресинаптического нейрона; t_{post} – время генерации импульса постсинаптическим (выходным) нейроном; $\Delta t = t_{pre} - t_{post}$ интервал между приходящим и генерирующимися импульсами, Δw – величина изменения синаптического веса; A_+ и A_- – максимальные положительное и отрицательное изменения синаптического веса, τ – коэффициент, определяющий диапазоны пре- и постсинаптических интервалов между импульсами, в течение которых возможно усиление или ослабление синапсов.

Для симуляции ИМНС была разработана программа на языке Python3 с использованием пакетов PyTorch и NumPy.

III. МОДЕЛИРОВАНИЕ ИМНС С УЧЕТОМ ОСОБЕННОСТЕЙ МЕМРИСТИВНЫХ КРОССБАР-МАССИВОВ

В первую очередь для перехода от абстрактной программной модели нейросети к модели с физически реалистичными характеристиками весовые коэффициенты синапсов были линейно преобразованы из условных единиц в единицы проводимости [Ом^{-1}] и пропорционально изменены параметры нейронов (табл. 1).

ТАБЛИЦА I. ПАРАМЕТРЫ АБСТРАКТНОЙ И УТОЧНЕННОЙ МОДЕЛИ

Параметр нейросети	Абстрактная модель, у.е.	Реалистичная модель
Минимальный/максимальный вес синапса	0 / 1	$5 \cdot 10^{-5} / 0.01 \text{ Ом}^{-1}$
Порог срабатывания нейрона	20	0.2 В
Приращение порога срабатывания нейрона	0.02	0.0002 В
Максимальное усиление/ослабление синапса	0.005 / 0.003	$5 \cdot 10^{-5} / 3 \cdot 10^{-5} \text{ Ом}^{-1}$

На рис. 2 приведена схема мемристового кроссбар-массива, модель которого представлена нами ранее в работе [9] и используется в итерационном алгоритме, позволяющем учесть нелинейность ВАХ мемристовых элементов и дискретность диапазона состояний проводимости при расчете токов и напряжений в кроссбар-массиве.

Разработанные модели ИМНС и мемристовых кроссбар-массивов, а также методики учета различных особенностей архитектуры и электронной компонентной базы, которые необходимо учитывать при аппаратном исполнении ИМНС, реализованы в виде пакета прикладных программ с помощью языка программирования Python3 и фреймворков PyTorch и NumPy.

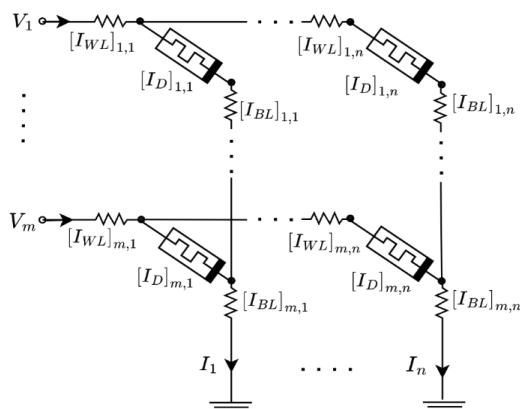


Рис. 2. Схема мемристового кроссбар-массива

IV. РЕЗУЛЬТАТЫ

Для апробации разработанных инструментов и оценки влияния особенностей функционирования мемристового кроссбар-массива на точность работы нейросети, были использованы экспериментальные ВАХ мемристовых структур на основе наноразмерной пленки TiOx , ВАХ которой приведены на рис. 1а.

ИМНС тестировалась в инференс-режиме (весовые коэффициенты были получены предварительно на абстрактной модели, а затем перенесены на физически-реалистичную модель) при амплитуде входных импульсов, равной напряжению чтения резистивных состояний мемристоров (0.5 В). Сопротивление элементов межсоединений используемой модели кроссбар-массива было выбрано 1 Ом.

Результаты компьютерного моделирования, показывают, что снижение точности распознавания изображений находится в диапазоне 7–11 % в случае учета сопротивления межсоединений, нелинейности ВАХ и дискретности набора состояний проводимости, по сравнению с изначальной точностью распознавания изображений абстрактной моделью. Наибольший вклад в снижение точности в случае инференс-режима вносит нелинейность ВАХ и изменение распределения выходных токов вдоль выходов строк кроссбар-массива, обусловленное падением напряжений на участках соединений. Результаты аналогового МВУ начинают ощутимо отклоняться от идеализированных программных расчетов, что вызывает снижение точности при распознавании изображений.

Для тестирования режима обучения ИМНС в аппаратном исполнении на архитектуре мемристового кроссбар-массива использовались аналогичные исходные параметры. С учетом того, что обучение напрямую затрагивает веса нейросети, изменяя их в течение всего процесса обучения, дискретность диапазона состояний проводимости стала сильно влиять на возможность обучения. На рис. 4 представлены результаты тестирования для различного количества уровней дискретизации, где показано, что без дополнительной корректировки параметров нейросети эффективность обучения резко падает при количестве состояний меньше 256.

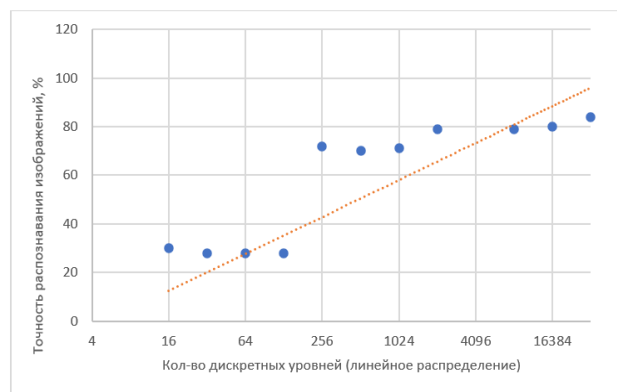


Рис. 3. Зависимость точности распознавания от количества дискретных уровней

Проблема возникает из-за размера шага между соседними уровнями, который увеличивается с уменьшением количества возможных состояний (при фиксированных границах всего диапазона). При правильно подобранных параметрах усиления и ослабления весов эту проблему можно частично решить, например, для 128 состояний возможно получить такой же результат обучения, как и при большем количестве состояний. Однако, при дальнейшем уменьшении количества уровней данный метод становится малоэффективен из-за слишком резкого изменения

весов, что делает обработку информации слишком «грубой».

Падения напряжений на межсоединениях также сильно нарушают динамику работы ИМНС в процессе обучения из-за неравномерного распределения выходных токов вдоль столбцов кроссбар-массива. Выходные токи на столбцах, находящихся на удалении от начала кроссбар-массива, куда подаются входные последовательности импульсов, могут оказаться на 5-10 порядков, а в случае увеличения размера кроссбар-массива на 10–30 порядков меньше, чем выходные токи на столбцах, расположенных ближе к началу кроссбар-массива. До определенной степени с этой проблемой можно бороться оптимизацией гиперпараметров нейросети, например, корректировать пороги срабатывания и их адаптивные составляющие, чтобы компенсировать неравномерность распределения приходящих сигналов на нейроны. Но выборочное изменение гиперпараметров в большинстве случаев вызывает нарушения баланса активности между нейронами, что также негативно влияет на успешность обучения нейросети. Моделирование аппаратного исполнения ИМНС в режиме обучения в случае подбора оптимальных гиперпараметров показало снижение точности распознавания изображений на 10–25 %, что согласуется с результатами, полученными в подобных работах.

V. ЗАКЛЮЧЕНИЕ

Полученные результаты показывают, что падения напряжения на элементах соединений кроссбар-массива, нелинейность ВАХ мемристивных элементов и увеличение дискретности диапазона состояний проводимости, которые необходимо учитывать при аппаратном исполнении ИМНС, в инференс-режиме на предобученных весах снижают точность распознавания изображений на 7–11 %.

При переходе в режим обучения, особенности архитектуры и мемристивных элементов могут приводить к тому, что сеть перестает обучаться, но оптимизация гиперпараметров ИМНС позволяет компенсировать возникающие негативные эффекты и уменьшить снижение точности распознавания

изображений после обучения до 10–25 % в сравнении с результатами, полученными на абстрактной программной модели, что говорит о необходимости учета данных факторов при проектировании нейроморфных модулей и разработке методов оптимизации гиперпараметров, а также об актуальности поиска принципиально новых решений для эффективного аппаратного исполнения алгоритмов ИМНС, что и является целью последующих исследований.

СПИСОК ЛИТЕРАТУРЫ

- [1] Thompson N., Greenewald K., Lee K., Manso G. The Computational Limits of Deep Learning, 2020. P. 46.
- [2] Chinae Manrique de Lara, Alejandro. (2019). Interpreting the High Energy Consumption of the Brain at Rest. *Proceedings*. 46. 6694. 10.3390/ecea-5-06694.
- [3] Ryndin, Eugeny & Andreeva, N.V. & Luchinin, Victor. (2022). Compact Model for Bipolar and Multilevel Resistive Switching in Metal-Oxide Memristors. *Micromachines*. 13. 98. 10.3390/mi13010098.
- [4] Peng Gu et al., "Technological exploration of RRAM crossbar array for matrix-vector multiplication," The 20th Asia and South Pacific Design Automation Conference, Chiba, Japan, 2015, pp. 106-111, doi: 10.1109/ASPDAC.2015.7058989.
- [5] S. A. Thomas, S. K. Vohra, R. Kumar, R. Sharma and D. M. Das, "Analysis of Parasitics on CMOS based Memristor Crossbar Array for Neuromorphic Systems," 2021 IEEE International Midwest Symposium on Circuits and Systems (MWSCAS), Lansing, MI, USA, 2021, pp. 309-312, doi: 10.1109/MWSCAS47672.2021.9531867.
- [6] L. Zhang, D. Borggreve, F. Vanselow and R. Brederlow, "Impact of Parasitic Wire Resistance on Accuracy and Size of Resistive Crossbars," 2021 IEEE International Symposium on Circuits and Systems (ISCAS), Daegu, Korea, 2021, pp. 1-5, doi: 10.1109/ISCAS51556.2021.9401279.
- [7] Deng L. (2012). The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6), 141–142.
- [8] Rukundo, Olivier & Cao, Hanqiang. (2012). Nearest Neighbor Value Interpolation. *International Journal of Advanced Computer Science and Applications*. 3. 25:30. 10.14569/IJACSA.2012.030405.
- [9] Dudkin A.P., Ryndin E.A. & Andreev, N.V. Modeling the Functional Features of a Memristive Crossbar Array in Neuromorphic Electronic Modules. *Russ Microelectron* 53, 554–566 (2024). <https://doi.org/10.1134/S1063739724600675>