НЕЙРОПРОЦЕССОРЫ ДЛЯ ИМПУЛЬСНЫХ НЕЙРОННЫХ СЕТЕЙ

Ларионов Д. А.[1,2], *Киселев М. В.*[1]

[1] Чувашский государственный университет имени И. Н. Ульянова, Чебоксары, Россия

[2] Частное учреждение «Цифрум», Госкорпорация «Росатом», Москва, Россия

Абстракт – В статье представлен обзор проектов, в рамках которых создаются или развиваются системы искусственного интеллекта, использующие импульсные нейронные сети. Фокус при рассмотрении сделан на особенностях аппаратного и программных дизайнов, проблематике на решение которой направлены проекты, их ключевых характеристиках и взгляде в будущее.

В статье рассмотрены следующие проекты: SpiNNaker, SpiNNaker2, DeepSouth, Tianji, Tianjic, TianjicX, ODIN, MorphIC, SPOON, ReckOn, DynapSEL, DynapCNN, DynapSE2, Speck, Xylo, Loihi, Loihi2, Spikey, BrainScaleS, BrainScaleS2, GrAI One, GrAI VIP, Akida, Innatera T1, TrueNorth и Алтай (AltAI).

Ключевые слова **– *нейроморфные вычисления, импульсные нейронные сети*.**

Обсудить материалы статьи и связаться с авторами возможно в telegram сообществе https://t.me/nrmairus.

Нейропроцессоры для импульсных нейронных сетей

Ларионов Д. А.[1,2], *Киселев М. В.*[1]

[1] Чувашский государственный университет имени И. Н. Ульянова, Чебоксары, Россия
[2] Частное учреждение «Цифрум», Госкорпорация «Росатом», Москва, Россия

Введение

В статье представлен обзор проектов, в рамках которых создаются или развиваются нейроморфные системы искусственного интеллекта (ИИ).

Согласно [1], под *нейроморфностью* понимается использование принципов организации и функционирования мозга. Под *системами искусственного интеллекта* понимается связка аппаратного обеспечения (hardware) и алгоритмов ИИ, которые на нем исполняются. На протяжении всей истории развития ИИ выигрывали именно те подходы, для которых находилось подходящее аппаратное обеспечение [2], поэтому рассматривать «железо» и алгоритмы важно именно в совокупности.

Идея нейроморфности однажды уже привела к появлению систем ИИ, имея в виду Джона фон Неймана [3] и Фрэнка Розенблатта [4], которые, создавая свои знаменитые концепты (архитектура фон Неймана и персептрон Розенблатта), пытались моделировать мозг. Однако, несмотря на значительный прогресс и впечатляющие результаты, современные системы ИИ по-прежнему далеки от своих биологических аналогов. Разрыв наблюдается в уровне энергопотребления, универсальности, адаптивности, масштабируемости [1].

Нейроморфные свойства

Сегодня в мире существуют десятки проектов, использующих идею нейроморфности для преодоления технологических барьеров в создании систем ИИ, однако каждый из проектов использует свой собственный уникальный набор нейроморфных свойств [1]:

- коннекционизм – дает обучение на данных;
- параллелизм – одновременное выполнение задач;
- асинхронность – отсутствие синхронизации, масштабируемость;
- импульсный характер передачи информации – простой протокол коммуникации, устойчивость к шуму;
- обучение на устройстве – непрерывное обучение;
- локальное обучение – снижение издержек на обучение, большие сети;
- разреженность – снижение издержек на перенос и обработку данных;
- аналоговые вычисления – эффективная аппаратная реализация, миниатюризация;
- вычисления в памяти – отсутствие издержек на перенос данных;
- квантизация весов – эффективная аппаратная реализация, миниатюризация;

Некоторые нейроморфные свойства уже повсеместно используются в существующих системах ИИ. Например, параллелизм и асинхронность, вычисления в памяти, импульсный характер передачи информации. Другие свойства, такие как локальное обучение, аналоговые вычисления – находятся на заключающих этапах научно-исследовательских проектов.

Импульсные нейронные сети

Одним из фундаментальных отличий между мозгом и системами ИИ, построенными на базе искусственных нейронных сетей (ИНС, artificial neural network, ANN) является импульсный характер передачи информации. Моделировать такой способ коммуникации позволяет аппарат импульсных нейронных сетей (ИмНС, spiking neural network, SNN).

В ИмНС вместо действительных чисел нейроны

обмениваются спайкаёми (spikes) – элементарными событиями, не имеющими никаких атрибутов, кроме времени их генерации. При этом передача спайка от нейрона к нейрону не происходит мгновенно, а требует некоторого времени, которое различно для разных пар нейронов [5]. Моделей нейронов в ИмНС в отличие от классических сетей существует множество. В задачах моделирования биологических структур используются более сложные модели (Ходжкина-Хаксли [7], Ижикевича [8]), в прикладных задачах чаще используется модель нейрона-интегратора с утечкой (leaky integrate-and-fire, LIF), либо интегратора с утечкой с адаптивным порогом (leaky integrate-and fire with adaptive threshold, LIFAT) [5] (Рис. 1).

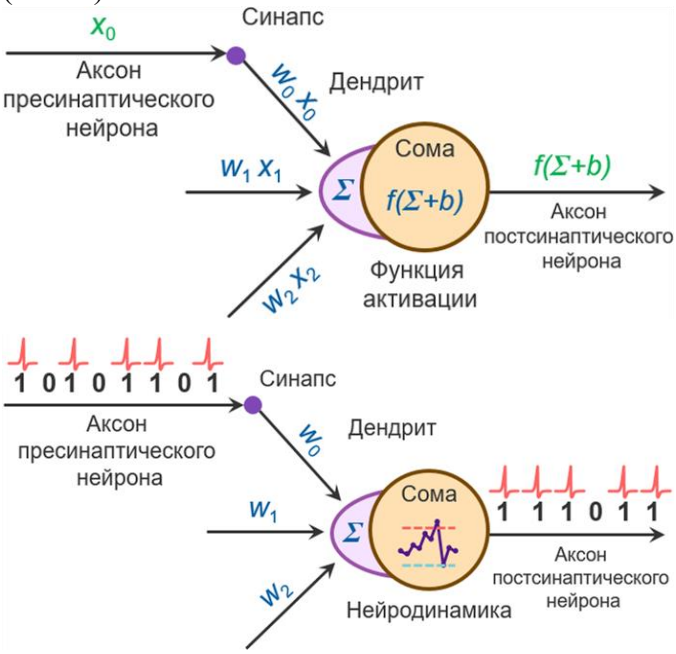

Рис. 1. Модель классического (сверху) и импульсного (снизу) нейронов.

Обучение ИмНС, как и классических нейронных сетей, как правило, основано на изменении весов межнейронных связей (синаптической пластичности). Однако в отличие от алгоритма обратного распространения ошибки (backpropagation) законы синаптической пластичности в ИмНС могут носить локальный характер [5]. Существует несколько подходов к обучению ИмНС:

- Обученная ИНС конвертируется в ИмНС с преобразованием активаций в частоты или времена спайков. Такой подход называют ANN2SNN.
- Backpropagation. Обучение происходит в домене ИмНС при помощи обратного распространения ошибки во времени.
- Локальное обучение. Изменение веса синапса зависит только от активности и состояния нейронов, соединенных данной синаптической связью. Такой подход совместим с полностью асинхронным характером функционирования ИмНС, а также является предпосылкой для непрерывного обучения и решения проблемы катастрофического забывания [6].

В контексте создания систем ИИ использование ИмНС позволяет обеспечить эффективный аппаратный дизайн, так как вычисление сети может быть сведено к аддитивным операциям (и редким умножениям). Более того, модель вычислений может быть полностью асинхронной, что является предпосылкой для масштабирования.

Таким образом, в части алгоритмов ИИ аппарат ИмНС позволяет «из коробки» работать с такими нейроморфными свойствами, как импульсный характер передачи информации, асинхронность и локальное обучение.

Другими важными преимуществами ИмНС являются:

- возможность работы с динамическим данными, так как время включается в явном виде в модель вычислений;
- простой асинхронный протокол коммуникации между нейронами, устойчивый к шуму;

К недостаткам ИмНС (в контексте создания систем ИИ) обычно относят [1]:

- низкую развитость программно-аппаратной экосистемы;
- превосходство классических алгоритмов ИИ в качестве решения прикладных задач;
- Открытые вопросы топологий и обучения ИмНС.

Обобщения ИмНС

Одним из популярных обобщений концепции ИмНС являются ИмНС с вариабельными спайками

(graded spikes). В таких сетях сигналы, параметризованы – обычно с ними связывается некоторая числовая величина. Такое обобщение делает ИмНС более близкими к традиционным нейросетям, в которых нейроны обмениваются действительными числами. В частности, это дает возможность реализовывать на нейроморфных вычислителях, поддерживающих вариабельные спайки, гибридные решения (в терминах ИНС и ИмНС).

Существуют также обобщения ИмНС, включающие генерализованные внесинаптические межнейронные взаимодействия. Однако эти модели обычно используются нейрофизиологами для моделирования работы мозга (взаимодействие между нейронами, включающее клетки глии, астроциты, гуморально-гормональное регулирование работы сети) [9], и не имеют прямого отношения к системам ИИ.

Обзор проектов

В данном обзоре рассмотрены проекты, которые используют аппарат ИмНС или его обобщения. Согласно [1], к классу нейроморфных систем ИИ также можно отнести множество других проектов, которые не используют аппарат ИмНС, однако они не включены в обзор, так как, по мнению авторов, именно ИмНС являются наиболее перспективным в области нейроморфных вычислений алгоритмическим подходом.

Такие перспективные направления как мемристивные вычисления [40], нейроморфная фотоника [41], резервуарные вычисления (reservoir computing) [42], проволочные сети (nanowire network) [43], квантовые импульсные сети [44], в рамках которых также используется аппарат ИмНС для создания систем ИИ, не включены в обзор из-за недостаточно высокого уровня зрелости решений на их основе (по сравнению с нейропроцессорами).

SpiNNaker, SpiNNaker2

Проект SpiNNaker создан с целью моделирования биологических структур мозга, допускает применение в робототехнике.

Основная идея – создание вычислительной системы на базе большого числа ARM (advanced RISC (reduced instruction set computer) machine) процессоров с большим объемом статической памяти (static random access memory, SRAM), на которых реализуется минимально достаточный набор инструкций для работы ИмНС с полноценным обучением. За счет асинхронности такая система может бесконечно масштабироваться [10].

Это свойство обеспечило в 2019 г. превосходство SpiNNaker над всеми другими аппаратными платформами в задаче моделирования работы 1 мм^2 кортикальной колонки [13]. Результатов работ по моделированию большего размера до сих пор не представлено.

Основные вехи проекта:

- **2005** Старт исследований в University of Manchester (Англия) под руководством *Steve Furber*
- **2013** Поддержка от Human Brain Project (HBP);
- **2016** Вычислительная система SpiNNaker (130 нм) для выполнения ИмНС из 500 тыс. ARM процессоров с поддержкой фреймворка PyNN [12];
- **2018** Масштаб 1 млн. ARM процессоров;
- **2019** Бенчмарк выполнения 1 мм^2 кортикальной колонки в реальном времени 77 тыс. нейронов, 285 тыс. синапсов с 0,1 мс интервалом [13];
- **2021** Запущен SpiNNaker2 (22 нм, 19 МБ SRAM) в сотрудничестве с TU Dresden (Германия). Добавлена поддержка ИНС через частотное кодирование, целый пласт акселераторов численных операций (exp, log, random, mac, conv2d), динамическое управление питанием [11];
- **2023** Более 100 систем SpiNNaker2 используются в лабораториях по всему миру, включая США, Японию, Австралию и Новую Зеландию;
- **2024** (анонс) Масштаб 5,2 млн. ARM процессоров.

DeepSouth

Проект DeepSouth использует схожую со SpiNNaker идею создания за счет асинхронности

бесконечно наращиваемой вычислительной системы для ИмНС, только вместо ARM-процессоров используются программируемые логические интегральные схемы (field-programmable gate array, FPGA).

Использование FGPA позволяет (за счет возможности переконфигурирования) менять за время исследований не только модели нейронов, пластичности синапсов, топологии сетей, стратегии маршрутизации, но и управлять квантизацией, балансом выделения памяти и другими аспектами, важными с точки зрения аппаратного дизайна.

Основные вехи проекта:

- **2023** На конференции ICNS анонсирована работа над проектом. Исследования ведутся в Western Sydney University (Австралия) под руководством *André van Schaik*;
- **апрель 2024** (анонс) Планировался запуск первой версии, но до сих пор не состоялся.

Tianji, Tianjic, TianjicX

Целью проекта является создание платформы для исследования сильного искусственного интеллекта (artificial general intelligence, AGI). Для этого исследователи строят унифицированную и гибридную (в терминах ИНС и ИмНС) систему ИИ.

В рамках проекта созданы процессоры, способные эффективно выполнять в режиме применения (inference) большие гибридные сети, за счет переиспользования одних и тех же участков схем для разных типов нейросетей. Гибридными сети могут быть не только в терминах слоев, но и в терминах отдельных нейронов, позволяя, например, создавать нейроны, которые на вход принимают спайки, а на выход возвращают действительные числа.

Представив в 2019 г. систему управления велосипедом на базе одного чипа Tianjic [15], исследователи сфокусировались на решениях для робототехники и столкнулись с проблемой согласованности работы нейросетевых модулей друг с другом [16], на решение которой направлены усилия настоящего времени [18].

Основные вехи проекта:

- **2013** Старт исследований в Tsinghua University (Китай) под руководством *Luping Shi*;
- **2015** Процессор Tianji (100 нм, 6 ядер, 256 КБ SRAM на ядро). Представлен архитектурный подход, объединяющий ИНС и ИмНС в рамках одной платформы. Не оптимизирован по всем параметрам, только классические ИНС и LIF нейроны [14];
- **2019** Процессор Tianjic (28 нм, 156 ядер, 40 тыс. нейронов, 10 млн. синапсов, 22 КБ SRAM на ядро, 6,1 мВт для ИНС и 5,5 мВт для ИмНС на ядро). Продемонстрирована система автономного управления велосипедом на базе одного чипа (обнаружение объектов, слежение, голосовое управление, обход препятствий, контроль равновесия), которая сочетала работу множества разных нейросетевых архитектур, включая ИмНС [15];
- **2022** Процессор TianjicX (28 нм, 160 тыс. нейронов, 20 млн. синапсов, 22,5 МБ SRAM), ориентированный на применение в робототехнике (низкое энергопотребление, низкая задержка, мультимодальный многозадачный параллелизм) [16]. Вместе с чипом представлена платформа для создания 4-лапых роботов и система Tianjicat (кошка, преследующая мышь);
- **2023** Масштабируемые архитектуры Tianjic card embedded systems и Tianjic board cloud servers [17];
- **2023** В статье «Coherence in Intelligent Systems» [18] выдвигается тезис о том, что именно согласованность работы нейросетевых модулей является мерой AGI.

ODIN, MorphIC, SPOON, ReckOn

Семейство процессоров для ИмНС с обучением. Проект развивается под руководством Шарлоты Френкель (*Charlotte Frenkel*) как исследовательская платформа для апробации нейроморфных подходов на конечных устройствах (at the edge).

Основная идея проекта ODIN состояла в следовании принципу «от биологии» [19], т.е. попытке учета максимально правдоподобных моделей нейронов и пластичности. На 2019 г. ODIN

за счет высокой плотности размещения синапсов показывал лучшие затраты энергии на одну синаптическую операцию и позволял запускать 20 нейронов Ижикевича [20].

В проекте ReckOn исследователи поставили главным принцип «от проблемы» [19], т.е. учет только достаточного набора нейроморфных свойств для решения прикладной задачи. Так в ReckOn упростились модели нейронов, а обучение перестало быть локальным [23].

Процессоры ODIN и ReckOn доступны в open-source (github.com/ChFrenkel) и могут быть прототипированы на FPGA. Доступна также облегченная версия (с LIF нейронами без пластичности) tinyODIN.

Основные вехи:

- **2016** Старт исследований в UCLouvain (Бельгия) под руководством *Charlotte Frenkel*;
- **2019** Open-source процессор ODIN (28 нм, 256 нейронов, 64 тыс. синапсов, локальное обучение на устройстве). 12,7 pJ/SOP (пикоджоулей на синаптическую операцию), 20 нейронов Ижикевича с локальным обучением (spike-driven synaptic plasticity) [20];
- **2019** Процессор MorphIC (65 нм, 2 тыс. нейронов, 2 млн. синапсов). Позволяет масштабировать ядра ODIN (local crossbar, inter-core tree-based and inter-chip mesh-based routing) [21];
- **2020** Старт исследований в ETH Zurich (Германия);
- **2020** Процессор SPOON (28 нм) для событийного зрения [22];
- **2022** Старт исследований в Delft University of Technology (Нидерланды);
- **2022** Open-source процессор ReckOn. Использует градиентный метод обучения (e-prop). Не использует внешнюю динамическую память с произвольным доступом (dynamic random access memory, DRAM). Потребление менее 50 мкВт [23].

Innatera T1

Коммерческий процессор на базе ИмНС для энергоэффективной непрерывной (always-on) обработки сенсорной информации для интернета вещей (internet of things, IoT) и носимых устройств. Не поддерживает обучение на устройстве.

Позволяет с использованием проприетарного фреймворка Talamo создавать и обучать ИмНС в задачах классификации паттернов и запускать обученные ИмНС на T1, достигая «из коробки»:

- классификация аудио сцен: 90%, 0,4 мВт, 301 мс;
- распознавание жестов: 97,5%, 0,4 мВт, 0,6 мс;
- детектирование присутствия человека: 91%, 0,4 мВт, <1 мс;
- распознавание звука: >95%, 0,72 мВт, 1 мс.

Основные вехи проекта:

- **2018** Образована компания Innatera как spin-off Delft University of Technology (Нидерланды);
- **2023** Представлен фреймворк Talamo (на базе PyTorch) для разработки и обучения ИмНС на CPU/GPU с использованием алгоритма backpropagation through time (BPTT), а также применения обученных ИмНС на совместимых платформах;
- **2024** Процессор T1 (~1 мВт, ~500 нейронов, 6 тыс. синапсов, 384 КБ SRAM), который включает в себя модуль выполнения в режиме применения ИмНС и процессорное ядро на базе RISC-V.

DynapSEL, DynapCNN, DynapSE2, Speck, Xylo

Семейство процессоров для ИмНС от компании SynSence. Данное семейство отличает большое разнообразие как в архитектурных подходах, так и в спектре решаемых задачах.

Первоначально процессоры SynSence развивались вокруг запатентованной идеи маршрутизации, основанной на автоматическом выборе баланса между широковещательными пакетами (broadcast) и коммуникацией по принципу каждый с каждым (point-to-point) (dynamic neurormorphic asynchronous processors,

DYNAP) [24]. Данная линейка содержит как исследовательские процессоры, созданные для изучения биологических структур и процессов передачи информации в мозгу, так и коммерческий процессор для задач компьютерного зрения DynapCNN.

Объединив DynapCNN с событийной камерой (dynamic vision sensor, DVS), исследователи представили интеллектуальный сенсор Speck, вокруг которого в 2024 г. SynSence объединились с компанией iniVation (производитель DVS).

SynSence также является разработчиком фреймворков sinabs (github.com/synsense/sinabs), Rockpool (github.com/synsense/rockpool), Samna и библиотеки для работы с событийными датасетами Tonic (github.com/neuromorphs/tonic).

Основные вехи:

- **2017** Образована компания SynSense как spin-off University of Zurich and ETH Zurich (Германия) под руководством *Giacomo Indiveri*;
- **2018** Процессор DynapSEL (self learning) для симуляции биологических структур с помощью ИмНС. Аналоговая реализация нейронов и цифровая коммуникация, обучение на устройстве (80 тыс. синапсов из которых 8 тыс. пластичны);
- **2019** Процессор DynapCNN (28 нм, 1 млн. нейронов с ReLU (rectified linear unit) активацией, до 9 сверточных слоев, до 16 классов, ~5 мВт) для инференса ИмНС в задачах зрения. ANN2SNN при помощи open-source фреймворка sinabs (на базе PyTorch);
- **2019** Представлена система зрения Speck в форме SoC (system on a chip), включающая DVS и DynapCNN;
- **2021** Процессор DynapSE2 (1024 аналоговых нейрона, 64 тыс. синапсов на нейрон, ~5 мВт) для инференса ИмНС, позволяющий выполнять сети прямого распространения, рекуррентные и резервуарные сети. Аналоговая реализация нейронов и цифровая коммуникация. Поддерживает фреймворки Rockpool, Samna, NEST, Brian2 [25];
- **2022** Процессор для работы с аудио и сигналами биосенсоров Xylo (28 нм, 1000 LIF нейронов, 278 тыс. синапсов, ~5 мВт, 8 бит на вес). ANN2SNN при помощи open-source фреймворка Rockpool (на базе JAX). Позволяет задавать индивидуальные временные константы синапсов и мембран, пороги и смещения для каждого нейрона, а также использовать рекуррентные связи и остаточные соединения (residual connections) [26];
- **2023** (анонс) DynapCNN2 для работы с 3D.

Loihi, Loihi2

Цифровой процессор для ИмНС с полноценным обучением. Разрабатывается компанией Intel в международном консорциуме Intel Neuromorphic Research Community (INRC). Начиная со второго поколения на чипе реализована поддержка обобщенной модели ИмНС, в которой спайки могут иметь значение до 32 бит (graded spikes), модель нейрона программируема, введены дополнительные механизмы пластичности.

Первые два поколения проекта Loihi являются исследовательскими проектами, направленными на поиск оптимальных решений для промышленной версии.

Проект Loihi реализуется в Intel's Neuromorphic Computing Lab под руководством *Mike Davis*.

Основные вехи проекта:

- **2018** Процессор Loihi для ИмНС с обучением на чипе (14 нм, 128 ядер и 3 CPU, 128 тыс. нейронов, 128 млн. синапсов). Конфигурируемая модель LIFAT нейрона. Доступны как градиентные методы обучения (backpropagation through time, BPTT), так и локальные (spike-timing dependent plasticity, STDP) [27];
- **2020** Продемонстрирована возможность локального обучения на устройстве (on-chip local learning) в задаче классификации запахов [28];
- **2021** Процессор Loihi2 (7 нм, 128 ядер и 6 CPU, 1 млн. нейронов, 120 млн. синапсов). Поддерживает вариабельные спайки 32 бита, локальные широковещательные пакеты (local broadcasts), обучение с

помощью алгоритма spike layer error reassignment in time (SLAYER), трехмерное масштабирование. Программируемая модель нейрона, поддержка трехфакторной пластичности для задач обучения с подкреплением (reinforcement learning, RL) [29];

- **2021** Open-source фреймворк LAVA для CPU, GPU и проприетарный модуль для Loihi2;
- **2024** Нейрокомпьютер Hala Point (1152 чипа Loihi2, 1,15 млрд. нейронов, 128 млрд. синапсов, 2600 Вт, 30 POPS (пета операций в секунду)) на базе Loihi2.

Spikey, BrainScaleS, BrainScaleS2

Семейство процессоров, которые используют идею аналоговых вычислений на основе RC-контуров (резистор и конденсатор). Проект создан с целью исследования процессов передачи информации в мозгу для создания прикладных решений в области робототехники.

Главной особенностью проекта является использование аналоговых вычислений для моделирования мембранного потенциала импульсного нейрона (при этом коммуникация остается цифровой). В первой версии BrainScaleS (BSS) ИмНС возможно было использовать только в режиме применения (inference).

Во втором поколении BSS2 за счет добавления дополнительных цифровых процессоров появилась поддержка классических ИНС и стало доступно обучение на устройстве. В том числе предложена концепция структурной пластичности, когда существующие связи переназначаются для других пар нейронов.

Основные вехи проекта:

- **1995** Старт исследований в Kirchhoff Institute of Physics, Heidelberg University (Германия) под руководством *Karlheinz Meier* (Карлхайнц Майер);
- **2004** Процессор Spikey для ИмНС с локальным обучением на базе STDP с аналоговыми элементами (RC-контуры);
- **2011-2015** Реализован проект по исследованию обработки информации в мозгу BrainScaleS [30]. Представлена интегральная схема специального назначения (application-specific integrated circuit, ASIC) (180 нм, 512 нейронов, 114 тыс. синапсов) с аналоговыми нейроядрами (но с цифровой коммуникацией между ними) для выполнения ИмНС без обучения. На базе процессора построена самая большая аналоговая вычислительная система BrainScaleS1 (19 тыс. нейронов, 1,4 млн. синапсов);
- **2013-2023** Реализован проект EBRAINS в рамках HBP. Представлена облачная система на базе процессора BrainScaleS2 [31]. С 2018 проектом руководит *Johannes Schemmel* (Йоханнес Шеммель);
- **2020** Процессор BSS2 (65 нм, 10 pJ/SOP, 512 нейронов по 256 синапсов, 2 CPU, поддержка обучения) сочетает как аналоговые ядра, так и CPU. Может работать с гибридными сетями (как ИНС, так и ИмНС), модель нейрона - Adaptive Exponential Integrate and Fire (AdEx). Позволяет полноценное обучение, в том числе за счет структурной пластичности. Поддерживает фреймворки PyNN и hxtorch (на базе PyTorch);
- **2023** Фреймворк jaxsnn на основе JAX, совместимый с аппаратной платформой BSS2, позволяющий реализовать обучение на устройстве с помощью алгоритма e-prop [32].

GrAI One, GrAI VIP

Семейство процессоров компании GrAI Matter Labs, которые используют свойство разреженности потоков данных для ускорения инференса как классических ИНС, так и ИмНС. В 2023 г. GrAI Matter поглощена Snapchat для использования в продукте Spectacles (умные очки).

Основная идея – ускорение вычислений за счет использования высокой скоррелированности фреймов (например, кадров) в задачах обработки потоковых данных (аудио, видео). Если активация нейрона не сильно изменилась по сравнению с прошлым тактом, то можно не посылать ее повторно. Такой подход позволяет резко снизить

количество синаптических операций [37].

Основные вехи проекта:

- **2016** Компания GrAI Matter Labs получила грант DARPA на коммерциализацию исследований Vision Institute Paris (Франция);
- **2019** Процессор для инференса в задачах с видео фреймами GrAI One [37]. Поддерживает как ИмНС, так и классические ИНС. Совместим с фреймворком Keras;
- **2020** Представлена архитектура Neuronflow и событийная модель вычислений SparNet для работы с высокоскоррелированными потоковыми данными, которая объединяет архитектурный подход Dataflow с разреженными вычислениями [38];
- **2021** Процессор для инференса в задачах с видео и аудио потоковыми данными GrAI VIP (28 нм, 196 ядер, 200 тыс. нейронов, событийная модель вычислений SparNet). Продемонстрировано выполнение MobileNetv1-SSD на 30 fps при потреблении 184 мВт.

AKIDA

Проект Akida компании BrainChip представляет из себя конфигурируемую IP-платформу для широкого класса задач обработки потоковых данных. Для создания решений на базе платформы Akida используются проприетарный подход Temporal Event Based Neural Networks (TENN) и фреймфорк MetaTF (на базе Keras). BrainChip идет по пути создания инструментов для конвертации популярных нейросетевых архитектур (включая ИмНС, трансформеры и др.) в модель TENN для выполнения на Akida.

Основная идея подхода TENN в том, что обучение происходит на GPU с поддержкой пространственных и временных 2D сверток, после чего обученная сеть конвертируется в рекуррентную структуру, эффективно выполняющуюся на Akida за счет большого объема SRAM и обходных связей (skip connections) между нейронами.

В первом поколении (AKD1000) подход TENN еще не применялся, но уже была реализована поддержка «обучения на устройстве» с помощью запоминания комбинации активаций на последнем полносвязном слое классической сверточной сети (convolutional neural network, CNN), что является инженерным трюком.

В 2024 г. на базе платформы Akida создана конфигурация Pico, которая впервые на конечном устройстве продемонстрировала инференс большой языковой модели (large language model, LLM) (в демонстрации использовалась обрезанная версия Llama3b, преобразованная в рекуррентную структуру Mamba, что также является инженерным трюком).

Основные вехи проекта:

- **2020** Процессор AKD1000 (28 нм, 80 нейроядер, 1,2 млн. нейронов, 10 млрд. синапсов, 100 мкВт – 300 мВт) для инференса ИмНС, которые получены конвертацией CNN (обученных с помощью Keras) при помощи проприетарного фреймворка MetaTF [39];
- **2023** Второе поколение Akida2 в виде конфигурируемой IP платформы для задач обработки потоковых данных. Добавлена поддержка квантизации и длинные обходные связи (long range skip connections), поддержка Vision Transformers (ViT);
- **2023** Предложен подход к обучению применению нейронных сетей TENN. Анонсирована поддержка формата ONNX в MetaTF;
- **2023** Напечатана первая партия AKD1500 (22 нм).

TrueNorth

Проект TrueNorth компании IBM стал первым промышленным процессором для ИмНС и, по всей видимости, получил применение в военной промышленности.

TrueNorth использует ANN2SNN подход (т.е. не поддерживает обучение на устройстве), ограничен по количеству связей, разрядности весов и другим параметрам, однако на базе TrueNorth представлена первая в мире событийная система зрения на базе DVS, продемонстрированы прикладные применения в задачах стереозрения и

распознавания большого числа объектов на видео большого разрешения.

После завершения работ над TrueNorth команда проекта сфокусировалась на проекте NorthPole, который стал самой совершенной на 2024 г. системой ИИ в части энергоэффективности и эффективности расходования транзисторов на площадь пространства. Однако, в проекте NorthPole ИмНС и асинхронная логика не используются, что переводит его в класс «обычных» нейроускорителей.

Основные вехи проекта:

- **2008** Группа *Dharmendra Modha* в IBM Research начала сотрудничество с DARPA в рамках проекта SyNAPSE;
- **2014** Первый промышленный процессор для ИмНС - TrueNorth (28 нм, 4096 ядер, 1 млн. нейронов, 256 млн. синапсов, 0,1 Вт, 6000 фреймов/Вт). LIF нейроны с линейной утечкой. ANN2SNN, веса связей ограничены 2 битами, 256 связями на нейрон [32];
- **2017** Представлена первая в мире событийная система зрения на базе DVS камеры и TrueNorth (10 жестов, 96,5%, 0,18 Вт). Опубликован датасет Gesture Recognition [33];
- **2018** Представлена система на базе TrueNorth из 4 акселераторов по 16 процессоров (NS16e-4) и продемонстрирована в задаче распознавания объектов в видео высокого разрешения. Начаты работы над проектом NorthPole [34];
- **2023** Процессор для инференса классических ИНС NorthPole (12 нм, 256 ядер, 224 МБ SRAM). В отличие от TrueNorth не использует асинхронную логику и не поддерживает ИмНС. Может выполнять множество популярных архитектур (EfficientNet-b7, YOLO-v4, BERT). Превосходит все существующие архитектуры в части энергоэффективности и эффективности расходования транзисторов на площадь пространства [35];
- **2024** Представлена система NorthPole VPX board. Продемонстрирован инференс LLM (3 млрд. параметров, 28 356 токенов/с) на системе из 16 карт. Анонсирована работа с LLM размером до 80 млрд. параметров [36].

Алтай (AltAI)

Цифровой процессор для инференса ИмНС с неограниченным масштабированием. Подходит для обработки потоковых данных для задач IoT, робототехники, компьютерного зрения.

Проект развивается компанией Мотив-НТ в сотрудничестве с Лабораторией Касперского, Курчатовским институтом и другими центрами компетенций. Первые два поколения Алтая позволяют выполнение ИмНС в режиме применения. Полноценную поддержку обучения планируется добавить в третьем поколении, которое будет ориентировано на задачи RL.

Одним из ключевых архитектурных особенностей третьего поколения Алтая являются программируемые с помощью микрокода ядра, что открывает возможность обучения на устройстве, а также позволяет использовать Алтай без внешнего вычислительного оборудования (в режиме «без хоста»).

Разработка и обучение ИмНС для применения на Алтае реализована с помощью open-source фреймворка Kaspersky Neuromorphic Platform (KNP, github.com/KasperskyLab/knp), который в том числе содержит программный эмулятор Алтая.

Основные вехи проекта:

- **2015** Старт исследований по проектированию ASIC для выполнения ИмНС на базе Новосибирского государственного технического университета под руководством *Валерия Канглера*;
- **2020** Произведена первая партия процессоров для инференса ИмНС Алтай (28 нм, 256 ядер, 8 тыс. нейронов, 512 на ядро, 0,5 Вт, 4мВт на ядро, 2200 кадров/с). Разработан минимально достаточный инструментарий для конвертации обученных CNN в конфигурацию процессора. Разработан программный эмулятор Алтая [45];

- **2022** Анонсирована совместимая с Алтай, CPU и GPU платформа Kaspersky Neuromorphic Platform (KNP). KNP поддерживает PyNN, реализует как ANN2SNN, так и возможность работы с ИмНС напрямую. Представлено прикладное применение в задаче определения скорости вращения лопасти (зрение);
- **2023** Произведена вторая партия процессора Алтай с улучшенной технологией корпусирования. Представлено применение в задаче определения экстремальной частоты мерцания диода (light-emitting diode, LED);
- **2024** KNP в open-source. Представлено прикладное применение Алтая в задаче детекции большого числа объектов в видео высокого разрешения;

ЗАКЛЮЧЕНИЕ

Согласно мнению рыночных аналитиков 2024 г. был отмечен пиком интереса обращения к идее нейроморфности. Так в начале 2023 г. тренд «Neuromorphic computing» находился до пика кривой Garther Hype Cycle, а в конце 2024 г. он находится уже после пика, на спаде – перед выходом на плато продуктивности.

В настоящее время выделяется сразу несколько коммерческих проектов в области нейроморфных вычислений - проекты компаний BrainChip, Innatera, SynSence, которые предлагают энергоэффективные решения для вычислений на конечных устройствах, используя ANN2SNN. ANN2SNN самый популярный на сегодня подход, он также используется в Tianjic, TrueNorth и Алтай.

Важно также отметить, что энергоэффективность систем ИИ не является прерогативой одного лишь подхода, основанного на аппарате ИмНС, и может быть достигнута при помощи других инженерных подходов (нейроморфных и нет). Это демонстрирует созданный командой TrueNorth нейроморфный процессор NorthPole, который не использует ИмНС, но является самой энергоэффективной системой ИИ.

Некоторые проекты пошли по пути универсальности в части поддержки как ИНС, так и ИмНС. По-настоящему гибридные сети получили поддержку в Tianjic (исследователи убеждены, что это путь к AGI) и в BrainScaleS2. В проектах Loihi2, Akida и GrAI VIP используется идея дополнительной событийной модели вычислений для работы на чипе (LAVA, TENN, Neuronflow соответственно), в которую можно конвертировать как обычные, так и импульсные сети.

С проблематикой обучения на устройстве и локального обучения связаны либо проекты, направленные на изучение процессов обработки информации в мозгу (SpiNNaker2, ODIN, ReckOn, DynapSE2, BrainScaleS2), либо глобальные исследовательские проекты в области робототехники (Loihi2).

Идея аналоговых вычислений пока не получила большого распространения и утилизируется главным образом в части умножения матриц на кроссбарах (BrainScaleS2) для вычисления ИНС.

Благодаря гибкости и отлаженному процессу производства цифровых чипов нейроморфные процессоры, по-видимому, останутся цифровыми в ближайшие годы. Они будут основаны на архитектурах, где многие вычислительные ядра соединены цифровой шиной данных. По этой шине передаются пакеты для обмена информацией между ядрами. Идея хранения состояния и весов нейронов в памяти SRAM кажется временной из-за того, что память SRAM имеет низкую плотность и является энергозависимой. Последнее означает, что невозможно отключить целые ядра чипа с покоящимися нейронами и включать их только тогда, когда они нужны. Мы с нетерпением ждем новых типов памяти, которые решат эту проблему [1].

Ссылки

[1] Ivanov D, Chezhegov A, Kiselev M, Grunin A and Larionov D (2022) Neuromorphic artificial intelligence systems. Front. Neurosci. 16:959626. doi: 10.3389/fnins.2022.959626

[2] Fankle, J., and Carbin, M. (2018). The lottery ticket hypothesis: finding sparse, trainable neural networks. arXiv preprint arXiv:1803.03635. doi: 10.48550/arXiv.1803.03635

[3] J Neumann von (1958). The computer and the brain. Yale Univ Press, New Haven

[4] Rosenblatt, F. (1962) Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Spartan Books, Washington DC.

[5] Киселев М. В. Импульсные нейронные сети. Представление информации, обучение, память. Palmarium academic publishing, Рига, 2020.

[6] McCloskey, M. and Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In Psychology of learning and motivation, volume 24, pages 109–165. Elsevier.

[7] HODGKIN AL, HUXLEY AF. A quantitative description of membrane current and its application to conduction and excitation in nerve. J Physiol. 1952 Aug; 117(4):500-44. doi: 10.1113/jphysiol.1952.sp004764. PMID: 12991237; PMCID: PMC1392413.

[8] E. M. Izhikevich, "Simple model of spiking neurons," in IEEE Transactions on Neural Networks, vol. 14, no. 6, pp. 1569-1572, Nov. 2003, doi: 10.1109/TNN.2003.820440.

[9] S. Gordleeva et al., "Situation-Based Neuromorphic Memory in Spiking Neuron-Astrocyte Network," in IEEE Transactions on Neural Networks and Learning Systems, vol. 36, no. 1, pp. 881-895, Jan. 2025, doi: 10.1109/TNNLS.2023.3335450.

SpiNNaker

[10] Furber, S. B., Galluppi, F., Temple, S., and Plana, L. A. (2014). The spinnaker project. Proc. IEEE 102, 652–665. doi: 10.1109/JPROC.2014.2304638

[11] Höppner, S., Yan, Y., Dixius, A., Scholze, S., Partzsch, J., Stolba, M., et al. (2021). The SpiNNaker 2 processing element architecture for hybrid digital neuromorphic computing. arXiv preprint arXiv:2103.08392. doi: 10.48550/arXiv.2103.08392

[12] Mayr, C., Hoeppner, S., and Furber, S. (2019). SpiNNaker 2: a 10 million core processor system for brain simulation and machine learning. arXiv preprint arXiv:1911.02385. doi: 10.48550/arXiv.1911.02385

[13] Van Albada, S. J., Rowley, A. G., Senk, J., Hopkins, M., Schmidt, M., Stokes, A. B., et al. (2018). Performance comparison of the digital neuromorphic hardware spinnaker and the neural network simulation software nest for a full-scale cortical microcircuit model. Front. Neurosci. 12, 291. doi: 10.3389/fnins.2018.00291

Tianjic

[14] L. Shi et al., Development of a neuromorphic computing system, 2015 IEEE International Electron Devices Meeting (IEDM), Washington, DC, USA, 2015, pp. 4.3.1-4.3.4, doi: 10.1109/IEDM.2015.7409624.

[15] Jing Pei et al. “Towards artificial general intelligence with hybrid Tianjic chip architecture”. B: Nature 572.7767 (2019), c. 106—111.

[16] Ma S, Pei J, Zhang W, Wang G, Feng D, Yu F, Song C, Qu H, Ma C, Lu M, Liu F, Zhou W, Wu Y, Lin Y, Li H, Wang T, Song J, Liu X, Li G, Zhao R, Shi L. Neuromorphic computing chip with spatiotemporal elasticity for multi-intelligent-tasking robots. Sci Robot. 2022 Jun 15;7(67):eabk2948. doi: 10.1126/scirobotics.abk2948. Epub 2022 Jun 15. PMID: 35704609.

[17] Pei, J., Deng, L., Ma, C. et al. Multi-grained system integration for hybrid-paradigm brain-inspired computing. Sci. China Inf. Sci. 66, 142403 (2023). https://doi.org/10.1007/s11432-021-3510-6

[18] Zheng, H., Shi, L. (2023). Coherence in Intelligent Systems. In: Hammer, P., Alirezaie, M., Strannegård, C. (eds) Artificial General Intelligence. AGI 2023. Lecture Notes in Computer Science, vol 13921. Springer, Cham. https://doi.org/10.1007/978-3-031-33469-6_36

Frenkel

[19] C. Frenkel, D. Bol and G. Indiveri, "Bottom-up and top-down Approaches for the design of neuromorphic processing systems: Tradeoffs and synergies between natural and artificial intelligence," Proceedings of the IEEE, vol. 111, no. 6, pp. 623-652, June 2023. doi:10.1109/JPROC.2023.3273520

[20] C. Frenkel, M. Lefebvre, J.-D. Legat and D. Bol, "A 0.086-mm² 12.7-pJ/SOP 64k-synapse 256-neuron online-learning digital spiking neuromorphic processor in 28nm CMOS," IEEE Transactions on Biomedical Circuits and Systems, vol. 13, no. 1, pp. 145-158, February 2019. doi:10.1109/TBCAS.2018.2880425

[21] C. Frenkel, J.-D. Legat and D. Bol, "MorphIC: A 65-nm 738k-synapse/mm² quad-core binary-weight digital neuromorphic processor with stochastic spike-driven online learning," IEEE Transactions on Biomedical Circuits and Systems, vol. 13, no. 5, pp. 999-1010, October 2019. doi:10.1109/TBCAS.2019.2928793

[22] C. Frenkel, J.-D. Legat and D. Bol, "A 28-nm convolutional neuromorphic processor enabling online learning with spike-based retinas," Proc. of IEEE International Symposium on Circuits and Systems (ISCAS), October 2020. doi:10.1109/ISCAS45731.2020.9180440
[23] C. Frenkel and G. Indiveri, "ReckOn: A 28nm sub-mm² task-agnostic spiking recurrent neural network processor enabling on-chip learning over second-long timescales," Proc. of IEEE International Solid-State Circuits Conference (ISSCC), Feb. 2022. doi:10.1109/ISSCC42614.2022.9731734
SynSence
[24] Moradi, S., Qiao, N., Stefanini, F., and Indiveri, G. (2017). A scalable multicore architecture with heterogeneous memory structures for dynamic neuromorphic asynchronous processors (DYNAPs). IEEE Trans. Biomed. Circ. Syst. 12, 106–122. doi: 10.1109/TBCAS.2017.2759700
[25] O Richter, C Wu, AM Whatley, G Köstinger, C Nielsen, N Qiao, G Indiveri, "DYNAP-SE2: a scalable multi-core dynamic neuromorphic asynchronous spiking neural network processor", Neuromorphic Computing and Engineering 4 (1), 014003
[26] Hannah Bos and Dylan Muir, "Sub-mW Neuromorphic SNN audio processing applications with Rockpool and Xylo", (2022), https://arxiv.org/abs/2208.12991
Intel
[27] M. Davies et al., "Loihi: A Neuromorphic Manycore Processor with On-Chip Learning," in IEEE Micro, vol. 38, no. 1, pp. 82-99, January/February 2018, doi: 10.1109/MM.2018.112130359.
[28] M. Davies et al., "Advancing Neuromorphic Computing With Loihi: A Survey of Results and Outlook," in Proceedings of the IEEE, vol. 109, no. 5, pp. 911-934, May 2021, doi: 10.1109/JPROC.2021.3067593.
[29] G. Orchard et al., "Efficient Neuromorphic Signal Processing with Loihi 2," 2021 IEEE Workshop on Signal Processing Systems (SiPS), Coimbra, Portugal, 2021, pp. 254-259, doi: 10.1109/SiPS52927.2021.00053.
BrainScaleS
[30] Schemmel, J., Brüderle, D., Grübl, A., Hock, M., Meier, K., & Millner, S. (2010). A wafer-scale neuromorphic hardware system for large-scale neural modeling. In Proceedings of the 2010 IEEE International Symposium on Circuits and Systems (ISCAS), (pp. 1947–1950).
[31] Pehle C, Billaudelle S, Cramer B, Kaiser J, Schreiber K, Stradmann Y, Weis J, Leibfried A, Müller E and Schemmel J (2022) The BrainScaleS-2 Accelerated Neuromorphic System With Hybrid Plasticity. Front. Neurosci. 16:795876. doi: 10.3389/fnins.2022.795876
[32] E. Müller, M. Althaus, E. Arnold, P. Spilger, C. Pehle and J. Schemmel, "jaxsnn: Event-driven Gradient Estimation for Analog Neuromorphic Hardware," 2024 Neuro Inspired Computational Elements Conference (NICE), La Jolla, CA, USA, 2024, pp. 1-6, doi: 10.1109/NICE61972.2024.10548709.
IBM
[32] Merolla, P. A., Arthur, J. V., Alvarez-Icaza, R., Cassidy, A. S., Sawada, J., Akopyan, F., et al. (2014). A million spiking-neuron integrated circuit with a scalable communication network and interface. Science 345, 668–673. doi: 10.1126/science.1254642
[33] Amir, A., Taba, B., Berg, D., Melano, T., McKinstry, J., Di Nolfo, C., et al. (2017). “A low power, fully event-based gesture recognition system,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, 7243–7252. doi: 10.1109/CVPR.2017.781
[34] DeBole, M. V., Taba, B., Amir, A., Akopyan, F., Andreopoulos, A., Risk, W. P., et al. (2019). Truenorth: accelerating from zero to 64 million neurons in 10 years. Computer 52, 20–29. doi: 10.1109/MC.2019.2903009
[35] Dharmendra S. Modha et al. Neural inference at the frontier of energy, space, and time. Science382, 329 - 335(2023). DOI:10.1126/science.adh1174
[36] Rathinakumar Appuswamy et al. Breakthrough low-latency, high-energy-efficiency LLM inference performance using NorthPole (2024), IEEE Conference on High Performance Exterme Computing (HPEC).
GrAI Matter
[37] Moreira, O., Yousefzadeh, A., Chersi, F., Kapoor, A., Zwartenkot, R.-J., Qiao, P., et al. (2020). “Neuronflow: a hybrid neuromorphic-dataflow processor architecture for AI workloads,” in 2020 2nd IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS) (IEEE),

Genova, Italy, 1–5. doi: 10.1109/AICAS48895.2020.9073999

[38] Khoei, M. A., Yousefzadeh, A., Pourtaherian, A., Moreira, O., and Tapson, J. (2020). “SparNet: sparse asynchronous neural network execution for energy efficient inference,” in 2020 2nd IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS) (IEEE), Genova, 256–260. doi: 10.1109/AICAS48895.2020.9073827

Akida

[39] Vanarse, A., Osseiran, A., Rassau, A., and van der Made, P. (2019). A hardware-deployable neuromorphic solution for encoding and classification of electronic nose data. Sensors 19, 4831. doi: 10.3390/s19224831

Алтай

[45] N.V. Grishanov et al., “Neuromorphic processor AltAI for energy-efficient computing,” Nanoindustry Russia, vol. 96, 2019, pp. 531-538.

Другие направления

[40] Li, Y., Wang, Z., Midya, R., Xia, Q., and Yang, J. J. (2018b). Review of memristor devices in neuromorphic computing: materials sciences and device challenges. J. Phys. D Appl. Phys. 51, 503002. doi: 10.1088/1361-6463/aade3f

[41] Yun-Jhu Lee, Mehmet Berkay On, Xian Xiao, Roberto Proietti, and S. J. Ben Yoo, Photonic spiking neural networks with event-driven femtojoule optoelectronic neurons based on Izhikevich-inspired model, Opt. Express 30, 19360-19389 (2022)

[42] Maas, W. Liquid State Machines: Motivation, Theory and Applications 275–296 (Imperial College Press, London, 2011).

[43] Zhu, R., Lilak, S., Loeffler, A. et al. Online dynamical learning and sequence memory with neuromorphic nanowire networks. Nat Commun 14, 6697 (2023). https://doi.org/10.1038/s41467-023-42470-5

[44] Kristensen, L.B., Degroote, M., Wittek, P. et al. An artificial spiking quantum neuron. npj Quantum Inf 7, 59 (2021). https://doi.org/10.1038/s41534-021-00381-7