Университет ИТМО, факультет программной инженерии и компьютерной техники
Двухнедельная отчётная работа по «Информатике»: аннотация к статье
Дата прошедшей лекции: 13.09.2022 Номер прошедшей лекции: 1 Дата сдачи: 27.09.2022

Выполнил Бардин П.А., № группы P3119, оценка _____
Фамилия И.О. студента не заполнять

Название статьи
The femtojoule promise of analog ai

ФИО автора статьи	**Дата публикации**	**Размер статьи**
Geoffrey W. Burr, Abu Sebastian, Takashi Ando, Wilfried Haensch	"20" ноября 2021 г.	3316 слов

Прямая полная ссылка на источник или сокращённая ссылка
https://spectrum.ieee.org/analog-ai

Теги, ключевые слова или словосочетания
Machine learning, CNN, DNN, Analog AI, von Neumann bottleneck, RAM, energy efficient AI

Перечень фактов, упомянутых в статье

1. Сложность нейронных сетей растет огромными темпами (до миллирарда раз больше операций на обучение модели за 8 лет), с чем текущие аппаратные решения уже не справляются.
2. Основная проблема – узкое место фон Неймана – низкая пропускная способность шины RAM-CPU/GPU в сравнении с объемом памяти и скоростью обработки данных в ней.
3. Предложенное решение позволяет производить вычисления «прямо в памяти» и основано на применении «решетчатых массивов» (crossbar arrays), которые представляют из себя 2 перпендикулярных набора проводников адресующих заключенные между слоями ячеки памяти по строкам и битам в них соответственно
4. На текущий момент используются технологии RRAM (резистивные RAM) и PCM (память с фазовыми переходами), они хорошо встраиваются в описанные массивы, а для считывания производится измерение сопротивления ячейки
5. В памяти слово представляет собой веса нейрона представленные проводимостью, при подаче тока на строку, на линиях битов получаем ток соответствующий весам, объединив выходы конденсатором и подавая на строку напряжение в течении времени соответствующего значению активации получаем сумму умножений весов и входных занчений в виде заряда
6. Расчетная производительность - 65 трлн. оп./с, при энергопотреблении в 100 раз меньше классических систем
7. Существуют проблемы с SNR, а также с тем, что расширение нейронной сети может быть произведено за счет большего числа ячеек, но не перезаписи весов из внешней памяти
8. Данная архитектура также предусматривает и ускорение обучения, так как в процессе обратного распространения достаточно подавать входное напряжение не на строки, а на стольбцы, интегрирую по току уже со строк, тем не менее обновление ячеек еще не эффективно

Позитивные следствия и достоинства описанной в статье технологии

1. Найдено новое напрвление, развитие которого позволит более эффективно масшабировать нейронные сети, значит будут доступны более сложные алгоритмы, в том числе и на более слабых устройствах / пользовательских устройствах
2. Данное решение представляет ценность как метод, применимый к различным технологиям, а значит имеет потенциал для дальнейшего развивития
3. Уход от классических решений приближает возможность повторения устройства мозга

Недостатки описанной в статье технологии

1. В конечном итоге масштабирование на большое количество нейронов упирается в скорость шины и устройства, обеспечивающих объединение нескольких вычислительных ячеек
2. Система требует высококачественных ADC, DAC для выполнения функций активации и операций между слоями, так как пока они реализуются в основном на процессорах
3. RRAM и PCM имеют технологические ограничения, в том числе связанные со сложностью записи данных, быстрым износом и высокой чуйствительностью к окружающим условиям
4. Проблемы с памятью усложняют процесс обучения и требуют использования решений для устранения шума, которые сказываются на эффективностималоэффективных