УДК 004.02+004.04+004.08
DOI 10.53403/9785951505071_2022_181

СПЕЦИАЛИЗИРОВАННЫЕ БОЛЬШИЕ ИНТЕГРАЛЬНЫЕ СХЕМЫ ДЛЯ РЕАЛИЗАЦИИ НЕЙРОСЕТЕВЫХ ВЫВОДОВ

Г. С. Елизаров, В. Н. Конотопцев, В. В. Корнеев

Научно-исследовательский институт «Квант», Москва

Представлены результаты экспериментального проектирования СБИС для реализации вывода сверточной нейросети SqueezeNet. Приведено сравнение прогнозируемых параметров этой СБИС с различными аппаратными платформами по производительности и производительности на ватт на тестовом наборе данных ImageNet.

Ключевые слова: сверточные нейросети, СБИС реализации вывода нейросети.

Для решения ряда важных научно-технических задач необходимо создание высокопроизводительных систем, использующих нейросетевые алгоритмы распознавания объектов и фона, навигации в пространстве. Ведущие университеты мира и корпорации, в том числе IBM и Intel, занимаются в научных и коммерческих целях разработкой эффективных нейросетевых парадигм и аппаратных архитектур для обучения и реализации выводов нейросетей. Высокая вычислительная сложность актуальных нейросетей привела к разработке библиотек для ускорения работы сетей на графических процессорах. Однако для высокопроизводительных моделей графических процессоров требуется мощность электропитания от 100 Вт, что делает проблемным их использование в автономных системах.

Для сокращения энергопотребления в процессе распознавания предлагается отделить процесс обучения и создавать нейросети, в которые сформированные ранее обучающие коэффициенты загружаются в виде максимально сокращенных массивов данных. Дополнительно сокращается трудоемкость отдельных операций путем замены операций с плавающей точкой на операции с фиксированной точкой.

В рамках НИР «Газель» (2015–2018 годы) по заказу Фонда перспективных исследований РФ была разработана технология быстрого проектирования СБИС. В качестве доказательства достижения поставленной цели был разработан проект специализированной СБИС Квант SN-2/2836, реализующей сверточную нейросеть (СНС) SqueezeNet [1]. Эта нейросеть выбрана в связи с тем, что в ней размер модели уменьшен приблизительно в 50 раз по сравнению с AlexNet, (AlexNet – 240 Мбайт [2]) при сохранении точности распознавания на том же уровне и применимости для решения разнообразных задач.

В СБИС реализованы все слои СНС SqueezeNet за исключением заключительного слоя softmax, вычислительная сложность которого невелика. Вычисления в СБИС Квант SN-2/2836 проводятся в формате чисел с фиксированной запятой, под целую часть отведено 13, под дробную – 11 разрядов. Точность вычислений при этом не хуже эталонного варианта реализации SqueezeNet с использованием арифметики с плавающей точкой. Вычислительная часть СБИС включает 2 синхронных конвейерных канала с общим управлением и общей памятью коэффициентов. Одновременно в обработке находятся до 4 кадров изображений (по два в каждом канале обработки) размером 224·224 пикселей. Полная обработка 2 кадров в СБИС занимает приблизительно $3{,}7 \cdot 10^6$ тактов работы. За каждый такт в каждом канале обработки выполняется 500 арифметических операций (умножение и сложение), всего в СБИС – 1 000 операций за такт.

Проект СБИС Квант SN-2/2836 был полностью промоделирован на ПЛИС Xilinx VC709 (XC7VX690T).

При изготовлении по технологии 28 нм на кристалле площадью 36 мм^2 прогнозируются следующие характеристики СБИС:

– частота работы – 1 200 МГц;

– мощность электропитания – от 12 до 14 Вт;

– производительность в числе кадров в секунду – 648 кадр/с;

– производительность в числе арифметических операций: 10^{12} оп/с.

В табл. 1 приведено сравнение эффективности реализации СНС SqueezeNet на различных аппаратных платформах на тестовом наборе данных ImageNet.

Приведенные данные демонстрируют высокую энергоэффективность прямой аппаратной реализации на СБИС этапа вывода (распознавания) для сверточных нейросетей SqueezeNet, а, также, возможность широкого масштабирования таких СБИС. Проектирование было осуществлено в директивный срок, предусмотренный разработанной в НИР «Газель» технологией.

Приведенные данные демонстрируют высокую энергоэффективность прямой аппаратной реализации на СБИС этапа вывода (распознавания) для сверточных нейросетей SqueezeNet – на обработку одного кадра затрачивается более чем на порядок меньше энергии, чем при использовании графических процессоров и более чем на два порядка меньше, чем в процессорах общего назначения.

Разработанная архитектура специализированной СБИС допускает широкое масштабирование вычислительных ядер, что позволяет обеспечивать требуемый уровень производительности для конкретных применений.

Проектирование СБИС Квант SN-2/2836а было осуществлено в директивный срок, предусмотренный разработанной в НИР «Газель» технологией.

Как результат активно протекающего процесса развития архитектуры нейронных сетей к настоящему времени разработаны нейросети, позволяющие еще более повысить достигнутые в проекте специализированного вычислителя Квант SN-2/2836 характеристики реализации нейросетевых выводов.

Среди направлений повышения производительности и энергоэффективности нейросетевых выводов [3] интенсивно развиваются бинарные нейросети с весами нейронов 1 или минус 1 и функциями активации с бинарными или целыми (int 32) значениями [4]. В бинарных сетях не используется энергозатратное умножение с плавающей точкой, существенно сокращается объем памяти для хранения весов.

Таблица 1

Эффективность реализации вывода СНС SqueezeNet на различных аппаратных платформах

Тип СБИС	CPU	GPU	SoC	FPGA	ASIC
Модель СБИС	Intel 7700k	Nvidia GTX 980Ti	Qualcomm Snapdragon 821	Xilinx VC709 (XC7VX690T)	Квант SN-2/2836
Технология, нм	14	28	14	28	28
Частота, ГГц	4,2/4,5	1/1,076	2,4	0,11	1,2
Вычислительные ядра, шт.	4	2816	4	2	2
Производительность, кадр/с	19,9	468,1	5,7	274,0	648,0
Мощность, Вт	104,5	203,5	4,6	27,7	14,0
Энергоэффективность, кадр/Дж	0,19	2,3	1,24	9,89	46,3

В табл. 2 приведено сравнение точности Top-1 верного распознавания нейросетей MPT-1/1 и MPT-1/32 с бинарными весами, сформированных и обученных по алгоритмам [4], и нейросетей с наилучшей достигнутой точностью с 32-разрядными весами на двух тестовых наборах данных CIFAR-10 и ImageNet.

Из табл. 2 видно, что объем памяти для хранения бинарных нейросетей меньше, чем требуется для нейросетей с 32-разрядными весами, но не в 32 раза, так как бинарная нейросеть может иметь большее число слоев и нейронов в слоях. Но отсутствие операций с плавающей точкой, делает вывод этих сетей более производительным и менее энергозатратным.

Таблица 2

Сравнение точности бинарных нейросетей МРТ и нейросетей с наилучшей достигнутой точностью с 32-разрядными весами на двух тестовых наборах данных CIFAR-10 и ImageNet

Нейросеть – разрядность весов/функции активации, тестовый набор	Память, Мбайт	Top-1, %
VGG-Small – 32/32, CIFAR-10	4,6	93,6
МРТ – 1/1, CIFAR-10	1,44	91,9
AlexNet – 32/32, ImageNet	240	57,2
SqueezeNet – 32/32, ImageNet	4,8	57,5
ResNet-34 – 32/32, ImageNet	21,8	73,27
МРТ – 1/1, ImageNet	19,3	52,07
МРТ – 1/32, ImageNet	13,7	74,03

Исходя из сегодняшнего состояния теории и практики нейросетевых вычислений разделение вывода и обучения представляется по-прежнему актуальным. Более того, появляются новые подходы к формированию и обучению нейросетей, например, на базе хеш-функций [5], выбора подсетей [4] и другие. Поэтому для формирования и обучения нейросетей требуются автоматизированные рабочие места с множеством алгоритмов формирования и обучения, а также наборов данных для достижения оптимального результата обучения.

Для вывода наиболее интересны нейросети с бинарными весами 1 или минус 1 и ассиметричной BPReLU (Biased parametric ReLU) функцией активации [6] с 1-, 2-, 32-разрядным выходом.

Представляется, что высокопроизводительная энергоэффективная реализация аппаратной платформы для выводов бинарных нейросетей может быть сделана на базе специализированных ПЛИС или СБИС, архитектура которых учитывает большую разреженность матриц весов и только два возможных значения их элементов: 1, минус 1. Поиск такой архитектуры – актуальная проблема.

Литература

1. Iandola F. et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size // arXiv:1602.07360v4 [cs. CV] 4 Nov 2016.

2. Alex K., Sutskever I., Hinton G.E. ImageNet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems // Lake Tahoe, NV, USA, 3–6 December 2012.

3. Корнеев В. В. Направления повышения производительности нейросетевых вычислений // Программная инженерия. 2020. Т. 11. № 1. С. 21–25.

4. Diffenderfer J., Kailkhura B. Multi-Prize Lottery Ticket Hypothesis: finding accurate binary neural networks by pruning a randomly weighted network // arXiv:2103.09377v1 [cs.LG] 17 Mar 2021 00:31:24 UTC (2,988 KB)

5. Chen B., Medini T., Farwell J., Gobriel S., Tai C., Shrivastava A. SLIDE : in defense of smart algorithms over hardware acceleration for large-scale deep learning systems // arXiv:1903.03129v2 [cs.DC] 1 Mar 2020

6. Zhang Y., Pan J., Liu X., Chen H., Chen D., Zhang Z. FracBNN: Accurate and FPGA-Efficient Binary Neural Networks with Fractional Activations // arXiv:2012.12206v1 [cs.LG] 22 Dec 2020 17:49:30 UTC (2,281 KB).

APPLICATION SPECIFIC INTEGRATED CIRCUITS FOR IMPLEMENTING NEURAL NETWORK INFERENCE

G. S. Elizarov, V. N. Konotoptsev, V. V. Korneev

Research Center “Kvant”, Moscow

The results of the experimental design of the VLSI circuit for the implementation of the inference of the convolutional neural network SqueezeNet are presented. The comparison of the predicted parameters of this VLSI circuit with various hardware platforms in terms of performance and performance per watt on the ImageNet test dataset is given.

Key words: convolutional neural networks, VLSI circuit implementation of neural network inference.