Аппаратная реализация нейронных сетей и нейро-нечетких преобразований на ПЛИС

А. М. Романов
Московский технологический университет (МИРЭА)
romanov@mirea.ru

***Аннотация*. В работе рассматривается вопрос аппаратной реализации нейронных сетей на базе ПЛИС. Предлагается подход, основанный на использовании сигма-дельта модулированных импульсных потоков для представления сигналов, позволяющий снизить ресурсоемкость как отдельных нейронов, так и межслойных соединений и обеспечить возможность полностью параллельной реализации нейронных сетей и нейро-нечетких преобразований на базе существующих микросхем ПЛИС. На основе предложенного подхода автором разрабатываются оригинальные ядра ПЛИС для реализации сетей прямого распространения, RBF-сетей и нейро-нечетких преобразований (ANFIS), эффективность работы которых демонстрируется в ходе натурных экспериментов.**

Ключевые слова: нейронные сети; нейро-нечеткое преобразование; ПЛИС; импульсные потоки; импульсные математические операции

I. Введение

Одним из ключевых достоинств нейронных сетей по сравнению с другими интеллектуальными технологиями является потенциальная простота их аппаратной реализации, которая позволяет не только существенно увеличить скорость расчета за счет использования массового параллелизма, но добиться повышения надежности за счет сохранения работоспособности при частичном или полном отказе отдельных нейронов или межслойных соединений. Наиболее перспективной платформой для аппаратной реализации нейронных сетей на сегодня являются программируемые логические интегральные схемы (ПЛИС). Они в отличие от микропроцессорных систем практически не накладывают на разработчика ограничений по структуре сети, функционалу нейрона и возможностям организации параллелизма, и в тоже время имеют существенно более низкую стоимость проектирования по сравнению с заказными сверхбольшими интегральными схемами (СБИС).

Главные сложности при реализации нейронных сетей на базе ПЛИС связаны с ограничениями этих микросхем по логической емкости и частоте. Так нейрон состоящий из двух умножителей и сумматора с ограничением потребует при реализации на широко распространённой ПЛИС Xilinx Artix XC7A12T более 600 логических ячеек, что составляет около 5% её ёмкости. Таким образом, на этой микросхеме можно разместить не более 20 таких простейших нейронов. Размер нейрона можно уменьшить за счет использования встроенных аппаратных умножителей. Однако, например, для рассмотренной выше микросхемы, их количество составляет 40 штук. То есть при полностью независимой реализации всех элементов нейронной сети ограничение в 20 нейронов останется неизменным.

Вторым ограничивающим фактором для реализации нейронных сетей на ПЛИС после логической емкости являются транспортные задержки в межслойных соединения. Попытка соединить два слоя сети, в каждом из которых по 10 нейронов, может привести к необходимости организации сотни соединений. А если учитывать, что каждое из них должно передавать импульсно-кодовую модуляцию разрядностью 12–16 бит, то окажется, что между различными элементами разных слоев на ПЛИС придётся организовать больше тысячи перекрестных связей. Из-за того, что по технологическим причинам количество слоев, содержащих внутренние межблочные соединения, на ПЛИС ограничено, крайне трудно организовать синхронную передачу информации не только между разными нейронами, но и между различными битами одной связи «точка-точка». Все это приводит к увеличению максимальных транспортных задержек и существенному падению максимальной частоты работы ПЛИС.

Описанные выше сложности прямо или косвенно связаны с применением ИКМ для представления сигналов в нейронной сети и традиционных алгоритмов их цифровой обработки. Поэтому для их преодоления требуется развитие новых принципов организации расчетов, представления сигналов и их передачи между элементами ПЛИС.

II. Импульсные математические операции

Перспективным путем снижения, как ресурсоемкости всех элементов нейронной сети, так транспортных задержек в межслойных соединениях является предельное уменьшение разрядности передаваемых и обрабатываемых данных. Добиться этого, не снижая точности представления данных, можно за счет кодирования их не при помощи импульсно-кодовой модуляции, а с использованием сигма-дельта модулированных потоков

Работа выполнена при финансовой поддержке Министерства образования и науки Российской Федерации (уникальный идентификатор ПНИЭР RFMEFI58016X0008)

[1–5]. Данный подход позволяет существенно снизить ресурсоемкость всем математических операций за счет сокращения размерности их операндов, а также упростить трассировку межслойных соединений, снизив транспортные задержки.

Представление сигналов в форме сигма-дельта модулированных импульсных потоков широко распространено в цифровой обработке сигналов (ЦОС) и в частности в аудиотехнике. Наиболее распространённым методом проведения операция над такими сигналами является их децимация в фильтре низких частот, выполнение всех необходимых преобразований в ИКМ форме и обратное преобразование в импульсных поток при помощи сигма-дельта модулятора [6, 7] (рис. 1, а)

Недостатками данного подхода являются: задержки вносимые процессом фильтрации, а также большая ресурсоемкость операций над многоразрядными ИКМ сигналами при их реализации на ПЛИС.

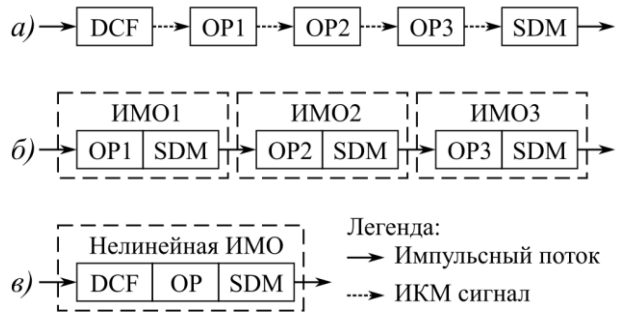

Рис. 1. Организация обработки сигналов (а) при помощи преобразования в ИКМ форму, (б) при помощи ИМО; (в) структура нелинейной ИМО. DCF – децемирующий фильтр, OP – математическая операция, SDM – сигма-дельта модулятор первого порядка

Альтернативой является прямая обработка сигма-дельта модулированных импульсных потоков при помощи импульсных математических операция (ИМО) [4, 5] (рис. 1, б,в). Первым отличием ИМО от обработки ИКМ сигналов являются полное отсутствие усредняющего фильтра на входе линейных операций и существенное уменьшение его периода усреднения для нелинейных операций. Это позволяет свести до минимума задержки, связанные с ЦОС, что крайне важно для задач управления, в которых присутствуют обратные связи. Вторым отличием ИМО является существенно меньшая разрядность операндов, которая для линейных ИМО составляет 1–2 бита, позволяя уменьшить ресурсоёмкость блоков ЦОС на ПЛИС до 14 раз [5]. Для нелинейных случаев (рис. 1, в), за счет сокращения периода усреднения децемирующего фильтра, удается обеспечить при использовании 5–7 битных операндов точность аналогичную 10–12 операциям над ИКМ.

В предыдущих работах для линейных и нелинейных ИМО автором сформулированы и доказаны теоремы и утверждения, позволяющие аналитически синтезировать операции, обладающие заданной точностью [4, 5, 8], а также разработаны программные продукты позволяющие проводить компьютерное моделирование и отладку ИМО [9]. Целью данной роботы является создание на базе ИМО базовых модулей ПЛИС, необходимых для реализации нейронных сетей основных типов, применяемых в задачах автоматического управления: сетей прямого распространения, RBF-сетей и нейро-нечеткие преобразований (ANFIS).

III. Состав модулей необходимых для реализации нейронных сетей

Все нейронные сети (рис. 2), перечисленные выше типов, состоят из ограниченного набора элементов: блоков масштабирования входов нейрона, сумматоров с ограничением, активационных функций различного типа и блоков нормализации.

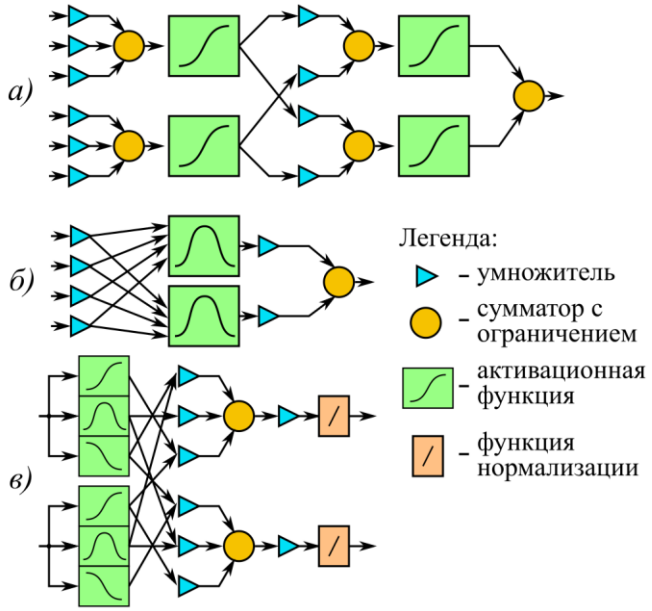

Рис. 2. Базовые элементы нейронных сетей различного типа: (а) – сетей прямого распространения, (б) – RBF-сетей, (в) – нейро-нечеткого преобразования ANFIS

Блоки масштабирования входов нейрона и сумматоры с ограничением являются по сути ранее описанными ИМО сложения импульсных потоков и умножения импульсного потока на число [4]. Реализация активационных функций типа «ограничение» при использование ИМО не требуется, так ограничение результата автоматически происходит на выходе сумматора импульсных потоков при его формировании в сигма-дельта модуляторе. Наилучшим способом реализации блока нормализации, используемого при подсчете взвешенной суммы в нейро-нечетких преобразованиях, будет использование ассоциативной памяти [8], так она гарантированно позволяет избежать каких-либо проблем, связанных с наличием сингулярности при нулевых значениях на входе.

В таблице представлены результаты сравнения ресурсоёмкости описанных выше базовых модулей

нейронных сетей на ПЛИС с их аналогами, реализованным с использованием традиционной обработки ИКМ.

Здесь и далее для сравнения ресурсоёмкости модулей используются результаты синтеза на ПЛИС Xilinx Artyx 7 XC7A100T при помощи пакета Xilinx ISE 14.7 (lin64).

ТАБЛИЦА I СРАВНЕНИЕ РЕСУРСОЁМКОСТИ РЕАЛИЗАЦИИ БАЗОВЫХ МОДУЛЕЙ НЕЙРОННЫХ СЕТЕЙ НА ПЛИС

Базовый модуль нейронной сети	**Ресурсоемкость реализации LUT6**	
	ИМО	***Обработка ИКМ***
Трехвходовой сумматор с ограничением	13	27
Умножитель на коэффициент 16 бит	49	297
Функция нормализации с точностью 10 бит	60	105

Данная работа посвящена в поиску подходов к реализации активационной функций для нейронных сетей различного типа.

IV. РЕАЛИЗАЦИЯ АКТИВАЦИОННЫХ ФУНКЦИЙ НЕЙРОНОВ НА БАЗЕ ИМО

A. Реализация активационной с использованием ассоциативной памяти

Ассоциативная память является универсальным способом реализации любой нелинейного операции над импульсными потоками [8]. Обобщённая структура блока активационной функции на базе ассоциативной памяти представлена на рис. 3.

Рис. 3. Обобщенная структура блока активационной функции на базе ассоциативной памяти. AF – усредняющие фильтры, MEM – статическая память, SDM – сигма-дельта модуляторы

Ресурсоёмкость реализации активационных функций на базе ассоциативной памяти напрямую зависит от объема входящей в её состав статической памяти и варьируется от 54 LUT6 при использовании 64 ячеек памяти по 16 бит до 476 LUT6 при использовании 2048 ячеек памяти по 16 бит. Объем памяти и ее разрядность выбираются разработчиком исходя из выбранного для конкретной задачи компромисса между требуемыми точностью и ресурсоёмкостью. Более подробная информация о зависимость между точностью и требованиями к статической памяти представлена в [8].

Ассоциативная память обеспечивает прогнозируемую ресурсоёмкость вне зависимости от типа и сложности аппроксимируемой нелинейности, что делает её хорошим инструментом для реализации RBF-сетей и нейро-нечетких преобразований.

B. Реализация активационных функция методом разложения в степенной ряд

Для сетей прямого распространения существует возможность получить активационную функцию меньшего размера, за счет аппроксимации ее нелинейности разложением в степенной ряд, реализованных при помощи ИМО.

Одной из наиболее распространённых активационных функций для сетей прямого распространения является th(x). Она достаточно хорошо аппроксимируется разложением в ряд Тейлора до 3 степени $th(x) = x — 0{,}25\,x^3$, которое обеспечивает точность порядка 7 бит. Структурная схема реализации данной аппроксимации на базе ИМО представлена на рис. 4.

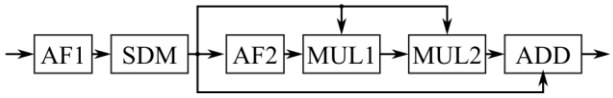

Рис. 4. Структурная схема блока реализации th(x) на базе разложения в степенной ряд, реализованного при помощи ИМО. AF – усредняющие фильтры, SDM – сигма-дельта модулятор, MUL – умножители импульсных потоков на коэффициент, ADD – двухвходовой сумматор импульсных потоков

Ресурсоемкость предложенного блока ИМО для аппроксимации th(x) составляет 50 LUT6, что на 10 LUT6 меньше реализации на базе ассоциативной памяти, обладающие сравнимой точностью.

Дальнейшее уменьшение ресурсоёмкости активационной функции методом разложения в ряд требует уменьшение количества используемых математических операций, и в первую очередь умножений. В данной работе была предложена новая активационной функция для сетей прямого распространяя $2x - x\,|x|$, оптимизированная с точки зрения необходимых для ее реализации ресурсов. По внешнему виду и характеристикам, она аналогична th(x): монотонна возрастающая, непрерывная и непрерывно дифференцируемая на интервале [-1:1], что делает возможным обучение сетей на её основе методом обратного распространения ошибки. Рассмотрение ее за пределами данного интервала при реализации на базе ИМО нецелесообразно, так как в ИМО ограничение происходит автоматически.

В тоже время эта активационная функция крайне эффективно реализуется на аппаратном уровне в виде $x+x - x\,|x|$, требуя помимо демодулятора всего один трехвходовой сумматор и один умножитель импульсного потока на константу. При этом функция модуля реализуется на базе подбитной инверсии, требуя дополнительно всего 7 LUT6.

Структурная схема реализации предложенной активационной функции на базе ПЛИС представлена на рис. 5.

Рис. 5. Структурная схема блока реализации $2x - x\,|x|$ на базе ИМО. AF – усредняющие фильтры, SDM – сигма-дельта модулятор, ABS – операция взятия модуля на основе побитной инверсии, MUL – умножитель импульсного потока на коэффициент, ADD – трёхвходовой сумматор импульсных потоков

Ресурсоемкость предложенного решения при точности аналогичной, рассмотренной выше реализации th(x), составляет 45 LUT, что на 10% меньше th(x) и на 25% меньше реализации этой же активационной функции на базе ассоциативной памяти.

C. Реализация синусоидальной активационной функции

Как было показано выше, разложение в степенной ряд с последующей реализацией при помощи ИМО позволяет получать крайне эффективные с точки зрения ресурсов структуры. Однако не трудно заметить, что при увеличении количества членов ряда потенциальная экономия ресурсов достаточно быстро сходит на нет. Это легко увидеть на примере синусоидально активационной функции. Для получения точности более 8 бит уже потребуется использовать разложение до пятой степени включительно (1).

$$\sin\frac{\pi x}{2}=\frac{\pi x}{2}-\frac{\pi^3 x^3}{8\cdot 3!}+\frac{\pi^5 x^5}{32\cdot 5!} \qquad (1)$$

Разложение (1) потребует для свое реализации как минимум 1 трехвходовой сумматор и 10 умножителей. Из таблицы видно, что их суммарная ресурсоёмкость составит около 500 LUT6. Однако, для получения аналогичной точности в 10 бит при помощи ассоциативной памяти потребуется 476 LUT6, что хоть и меньше, но все же сравнимо с разложением в ряд.

Для существенно сокращения ресурсоёмкости синусоидальной активационной функции в данной работе предложена новая аппроксимация (2).

$$\sin\frac{\pi x}{2}=16.49\left(\frac{-1+\sqrt{1+(0.5409x)^2}}{0.5(0.5409x)}-0.4456x\right) \qquad (2)$$

Аппроксимация (2) позволяет получить сравнимую с ассоциативной памятью и степенными рядами точность и при этом реализуется с меньшим количеством ресурсов ПЛИС. При помощи (2) количество блоков умножителей импульсных потоков можно сократить до 4 штук, а нелинейные преобразования извлечения корня и деления реализовать при помощи схемотенических решений предложенных в [1] на основе блока интегрирования с умножителем импульсных потоков в обратной связи.

Итоговая ресурсоемкость синусоидальной активационной функции на базе аппроксимации (2), обеспечивающей точность значащих 10 бит, составляет 160 LUT6, что более чем на 65% меньше, чем аналогичные решения на основе степенных рядов и ассоциативной памяти.

V. Выводы

В работе рассмотрен вопрос эффективной реализации нейронных сетей на базе ПЛИС. В качестве метода сокращения транспортных задержек в межслойных соединениях и снижения объема ресурсов, требуемых для реализации нейронов на ПЛИС, предложено использование импульсных математических операций. Эффективность предложенного метода продемонстрирована на базе часто используемых в нейронных сетях блоков умножения на константу и сложения с ограничением и нормализации.

Основная часть работы посвящена различным методам реализации активационных функций нейронов. В качестве универсального подхода предложено использование нелинейных импульсных математических операций, построенных на базе ассоциативной памяти. Для сетей прямого распространения предложена как высокоэффективная реализации активационной функции типа th(x), так и новая активационная функция $2x - x\,|x|$, оптимизированная с точки зрения аппаратных ресурсов, необходимых для ее реализации ресурсов. Для синусоидальной активационной функции предложена новая аппроксимация, позволяющая при реализации с использованием импульсных математических операций обеспечить сокращение ресурсоёмкости более чем на 65% по сравнению с аналогами. Все предложенные решения апробированы на ПЛИС Xilinx Artyx 7 XC7A100T.

В совокупности разработанные автором базовые модули ПЛИС позволяют реализовывать на ПЛИС все основные типа нейронных сетей, используемые в задачах управления: сети прямого распространения, RBF-сети и нейро-нечеткие преобразования.

Список литературы

[1] Fujisaka H. et al. Bit-stream signal processing and its application to communication systems //IEE Proceedings-Circuits, Devices and Systems. 2002. Т. 149. №. 3. С. 159-166.

[2] Ng C.W., Wong N., Ng T. S. Bit-stream adders and multipliers for tri-level sigma–delta modulators //IEEE Transactions on Circuits and Systems II: Express Briefs. 2007. Т. 54. №. 12. С. 1082-1086.

[3] Sadik A.Z., O'Shea P.J. Realization of ternary sigma-delta modulated arithmetic processing modules //EURASIP Journal on Advances in Signal Processing. 2009. Т. 2009. С. 8.

[4] Романов А.М. Анализ и синтез элементов устройств управления мехатронно-модульными системами на базе ПЛИС с использованием сигма-дельта модуляции // Естественные и технические науки. М.: Спутник+, 2013. № 6., с. 348-361.

[5] Романов А.М Развитие технологии сигма-дельта модуляции для создания в архитектуре плис ресурсоемких устройств управления мехатронно-модульными системами: Автореф. дис. ... канд. техн. наук / МИРЭА, 2014. 24 с.

[6] Maloberti F. Non conventional signal processing by the use of sigma delta technique: a tutorial introduction //Circuits and Systems, 1992. ISCAS'92. Proceedings., 1992 IEEE International Symposium on. IEEE, 1992. Т. 6. С. 2645-2648.

[7] Reiss J., Sandler M. Digital audio effects applied directly on a DSD bitstream //Proc. of the 7th Conference on Digital Audio Effects (DAFx'04). 2004. С. 5-8.

[8] Alexey R., Mikhail R. FPGA based implementation of content-addressed memory based on using direct sigma-delta bitstream processing //Young Researchers in Electrical and Electronic Engineering Conference (EIConRusNW), 2016 IEEE NW Russia. IEEE, 2016. С. 320-324.

[9] Romanov A., Bogdan S. Open source tools for model-based FPGA design //Control and Communications (SIBCON), 2015 International Siberian Conference on. IEEE, 2015. С. 1-6.