Реализация АРТ-1 классификатора на ПЛИС

О. И. Буренева[1], М. С. Прасад[2], Шивани Верма[2]
[1]*Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)*
[2]*Университет Эмити, Институт космической науки и технологий Эмити, Индия*
[1]OIBureneva@etu.ru

***Аннотация*. Один из вариантов создания эффективного аппаратного классификатора связан с использованием сетей, построенных на базе адаптивной резонансной теории (АРТ). АРТ сеть имеет регулярную структуру, что обеспечивает максимальное распараллеливание процессов обработки. В работе представлен вариант реализации АРТ классификатора, ориентированный на имплементацию в ПЛИС, показана архитектура, описаны решения, реализованные при описании элементов сети, приведены результаты моделирования, подтверждающие эффективность классификатора.**

Ключевые слова: адаптивная резонансная теория; нейронная сеть адаптивного резонанса; классификатор; аппаратная реализация нейросети; программируемые логические интегральные схемы

I. Введение

Одной из задач, наиболее часто решаемых искусственными нейронными сетями (ИНС), является задача распознавания, состоящая в соотнесении входного вектора с некоторым классом. Для решения этой задачи традиционно используются классификаторы, построенные на основе сети адаптивно резонансной теории (АРТ), а также различные модификации сетей Хопфилда.

Адаптивная резонансная теория была представлена в [1] и развивается до настоящего времени [2], и на ее базе разрабатывается разновидность нейронных АРТ сетей. В отличие от большинства существующих архитектур ИНС АРТ сети не предусматривают деления жизненного цикла на стадии обучения и непосредственного распознавания. Обучение АРТ сетей происходит в процессе функционирования путем сохранения «знаний», полученных на основе уже обработанных данных и сформированных шаблонов. АРТ сети решают проблему стабильности-пластичности: они остаются статичными при поступлении данных, не являющихся для них актуальными, и проявляют свойство пластичности, адаптируясь к значимым входным данным.

На основе АРТ разработаны различные классы сетей, используемых для решения задач предсказания, классификации и распознавания образов [3]. АРТ-1 – модель нейронной сети, являющаяся классификатором двоичных входных данных, обучающимся без учителя [4]. На ее базе построены сети АРТ-2 [5] и нейронная сеть Fuzzy APT [6]. АРТ-2 позволяет обрабатывать вектора, представленные вещественными числами, что обеспечивается введением в архитектуру сети специальных слоев, в которых выполняется нормализация входных сигналов. Сеть Fuzzy APT обрабатывает как двоичные, так и аналоговые входные сигналы, используя операторы, предназначенные для работы с нечеткими множествами. Также известны другие варианты сетей ART (Adaptive Resonance Theory): Gaussian ART, Hypersphere ART, Gaussian ARTMAP, Hypersphere ARTMAP, Fuzzy ARTMAP, TopoART и т. д. [7–9].

При отсутствии ограничений на потребление мощности и массогабаритные характеристики нейросетевых модулей в качестве базы для аппаратной реализации ИНС используются графические процессоры (GPU) [10]. Они обеспечивают параллельные вычисления, что позволяет максимально быстро выполнять преобразования, предусмотренные систолической архитектурой нейронных сетей. Распараллеливание вычислительных операций в ИНС можно обеспечить и при их реализации на базе программируемых логических интегральных схем (FPGA) [11], а также при изготовлении в виде заказных СБИС [12]. Кроме высокого быстродействия за счет естественного распараллеливания вычислений такие решения характеризуются низким энергопотреблением и минимальными массогабаритными параметрами. Эти особенности делают возможным размещение нейросетевых модулей непосредственно в периферийных устройствах, получая при этом ряд дополнительных преимуществ:

- минимизацию времени обработки данных за счет формирования результата в точке получения информации без использования каналов передачи;
- обеспечение конфиденциальности информации из-за отсутствия передач по каналам связи.

Целью данной работы являлось построение аппаратной модели классификатора на базе адаптивной резонансной теории, используемого для распознавания и кластеризации бинарных векторов и ориентированного на реализацию в базисе программируемых логических интегральных схем.

II. Сети АРТ

A. Архитектура сети АРТ-1

В соответствии с [4] сеть АРТ-1 состоит из двух слоев нейронов. Первый слой – входной или слой сравнения, количество нейронов в нем определяется размерностью входного двоичного вектора. Второй слой – выходной или распознающий, количество

Работа выполнена при поддержке Минобрнауки России в рамках соглашения № 075-15-2020-933 от 13.11.2020 г о предоставлении гранта в форме субсидий из федерального бюджета на осуществление государственной поддержки создания и развития научного центра мирового уровня «Павловский центр «Интегративная физиология – медицине, высокотехнологичному здравоохранению и технологиям стрессоустойчивости».

нейронов в нем зависти от количества кластеров, на которые классифицируются объекты. Нейроны первого и второго слоев связаны восходящими (от входных к выходным) и нисходящими (от выходных к входным) связями по принципу «каждый – с каждым». Для каждой связи определен вес:

- w_{ij} – вес восходящей связи i-го нейрона входного слоя к j-му нейрону выходного;
- tji – вес нисходящей связи j-го нейрона выходного слоя к i-му нейрону входного.

Веса характеризуют образ, который определяет выходной нейрон, поэтому они относятся к долговременной памяти АРТ сети. Распознаваемый образ, обрабатываемый во входном и выходном слое, относят к кратковременной памяти. Входной слой на основе поступившего входного сигнала генерирует сигналы с использованием весов восходящих связей w_{ij}. В выходном слое ищется нейрон с наиболее похожим образом. Выходной слой генерирует обратный сигнал во входной слой с использованием весов нисходящих связей t_{ji}. Во входном слое выполняется сравнение входного образа и образа, пришедшего от выходного слоя. Если образы с учетом критерия похожести совпадают, то считается, что в сети возник резонанс и образ классифицирован. С учетом классифицированного образа корректируются соответствующие ему веса. Если совпадение не выявлено, то запускается новый цикл поиска.

B. Функционирование сети АРТ-1

Входной двоичный вектор X = (x_1, x_2,. . . , x_n) поступает на слой сравнения. Для работы АРТ сети необходимо определить пороговое значение R_t, где $0 \le R_t \le 1$, характеризующее степень похожести входного вектора x на шаблон класса: чем больше значение порога, тем больше будет сходство входного вектора и образа из кластера, к которому будет отнесен входной вектор. В момент начала работы в выходном слое определен только один нейрон, его шаблон определяется следующими весами:

$$w_{i1}^{0} = \frac{L}{L-1+n}, \quad (1)$$

$$c_{1i}^{0} = 1, \quad (2)$$

где w – вес восходящей связи; c – вес нисходящей связи; i – индекс входного нейрона, $1 \le i \le n$; L – константа, со значением более 1, определяющая степень влияния нового входного образа на кратковременную память, рекомендованное значение L=2.

При подаче на вход АРТ сети входного вектора X для каждого нейрона выходного слоя вычисляется функция

$$y_j = \sum_{i=1}^{n} w_{ij}^{t} x_i, \; 1 \le j \le m$$

где m – количество выходных нейронов; t – шаг работы.

Из полученного набора значений y_j выбирается максимальное значение $y_k = \max_j y_j$. Если значение максимума зафиксировано у нескольких значений, то выбирается выходной нейрон с минимальным значением индекса.

Выбранный выходной нейрон y_k сравнивается со значением входного вектора на основании количественной меры сходства входного вектора с интегрированным образом кластера, который рассчитывается следующим образом:

$$R_k = \frac{\sum_{i=1}^{n} w_{ik} x_i}{\sum_{i=1}^{n} x_i}.$$

Если значение R_k равно или превышает пороговое значение R_t, то входной вектор относится к выходному кластеру k. В противном случае считается, что сеть выявила новый образ, который не похож ни на один из хранящихся в долговременной памяти. Если еще существуют выходные нейроны, для которых образ не зафиксирован, или есть возможность увеличить количество выходных нейронов, то будет сформирован нейрон для выделения нового кластера с новым образом. Определенное влияние на количество кластеров оказывает значение константы L: при больших ее значениях функция состояния для незафиксированного нейрона будет превышать функции для нейронов с уже хранящимися образами. В результате для новых значений входных векторов X будут создаваться новые классы, что приведет к увеличению количества кластеров (росту количества нейронов в выходном слое) и неэффективной классификации.

Если входной вектор отнесен к существующему кластеру, то пересчитываются веса связей соответствующего нейрона:

$$w_{i\,k}^{t+1} = \frac{L c_{ki}^{t} x_i}{L-1+\sum_{i=1}^{n} c_{ji}^{t} x_i}, \quad (3)$$

$$c_{k\,i}^{t+1} = c_{k\,i}^{t} x_i, \quad (4)$$

Если на основе входного вектора формируется новый кластер, то веса соответствующего выходного нейрона формируются по следующим выражениям:

$$w_{ik}^{t+1} = \frac{L x_i}{L-1+\sum_{i=1}^{n} x_i}, \quad (5)$$

$$c_{ki}^{t+1} = x_i. \quad (6)$$

III. Аппаратная реализация сети АРТ-1

Описание сети АРТ-1 выполнено на языке проектирования аппаратуры SystemVerilog, что обусловлено необходимостью работы с матрицами. В первую очередь это описание ориентировано на дальнейшую имплементацию сети в программируемые логические интегральные схемы, однако оно может быть адаптировано и для создания макроблока заказной микросхемы.

Принципиальное отличие аппаратной реализации сети АРТ состоит в том, что ее архитектура определяется на этапе описания, то есть количество кластеров заранее зафиксировано. На этапе проектирования сети в выходном слое будет создано необходимое количество нейронов, которые в результате работы нейронной сети, будут обучены на выявление определенных образов из входных векторов. В случае, когда классификация будет успешной, один из нейронов выходного слоя будет активизирован. Если же для входного вектора не будет найден образ, удовлетворяющий выбранному критерию похожести, ни один из выходных векторов не активизируется, но при этом и дополнительный нейрон в выходном слое создаваться не будет. Отсутствие

активного нейрона на выходе будет показателем того, что входной вектор не может быть классифицирован этой нейронной сетью.

Как было отмечено ранее, в АРТ сетях отсутствует разделение процессов обучения и работы. В случае аппаратной реализации, при уже известном количестве выходных нейронов можно воспользоваться программной моделью сети для начального определения весов восходящих и нисходящих связей. Эти веса могут быть загружены в регистры модуля сети по сигналу сброса, необходимого для приведения аппаратуры в исходное состояние. По этому сигналу будут обнуляться регистры кратковременной памяти сети и предустанавливаться регистры долговременной памяти. Если не выполнять предобучение сети, то необходимо в качестве первых входных векторов представлять сети вектора, соответствующие конкретным образам, которые и обеспечат формирование кластеров.

Еще одна особенность аппаратной реализации сети состоит в необходимости масштабирования весовых коэффициентов, поскольку на ПЛИС специальные средства для работы с вещественными значениями не предусмотрены. В рассматриваемом варианте было выполнено масштабирование: в регистрах, хранящих веса, выделены биты, представляющие целую и дробную части хранимых вещественных чисел.

При реализации сети потребуется создать две группы нейронов для формирования входного и выходного слоя соответственно. Реализация операции умножения предполагает использование встроенных умножителей, что приводит к необходимости использования микросхем класса FPGA, где такие умножители имеются. Дополнительно необходимо разработать блок управления, блок выявления нейрона с максимальным значением функции состояния и схему сброса. Описание архитектуры сети легко параметризуется, что позволяет оперативно изменять количество нейронов в разных слоях. Сложности с параметризацией возникают в модуле выявления максимального значения, причем в том случае, когда он имеет древовидную структуру, обеспечивающую максимальное быстродействие [13]. При последовательном выявлении максимума, например методом пузырьковой сортировки, проблема параметризации отсутствует.

Архитектура сети АРТ-показана на рис. 1. В состав сети входят нейроны слоя сравнения Z и нейроны слоя распознавания Y. Для управления работой сети введены блоки управления G_1 и G_2, которые определяют этап работы, а также нейрон сброса R, переводящий в активное состояние нейрон выходного слоя Y, показывающий результат и выключающий все остальные Y-нейроны. Веса связей, идущих от нейронов Z к нейронам Y рассчитываются по формулам 1, 3, 5, а веса обратных связей по формулам 2, 4, 6 соответственно. Возбуждающие связи: от входных элементов X к элементам Z и к нейронам R, G_1; от нейронов G_1, G_2 и R к нейронам слоев Z и Y. Тормозящие связи: от интерфейсных элементов X к R-нейрону; от Y-нейронов к управляющему блоку G_1.

Нейроны сравнения Z имеют три входа: бит входного вектора x_i; взвешенный выход нейронов второго слоя $y_j c_{ji}$, где y_j – выход j-го нейрона второго слоя, c_{ji} – весовой коэффициент нисходящей связи от j-го выходного нейрона к i-му входному; сигнал блока управления G_1, формирующийся при наличии на входе сети классифицируемого вектора и отсутствии результата на выходе слоя распознавания, что соответствует работе сети: $G_1 \sim (|Y)\ \&\ (|X)$.

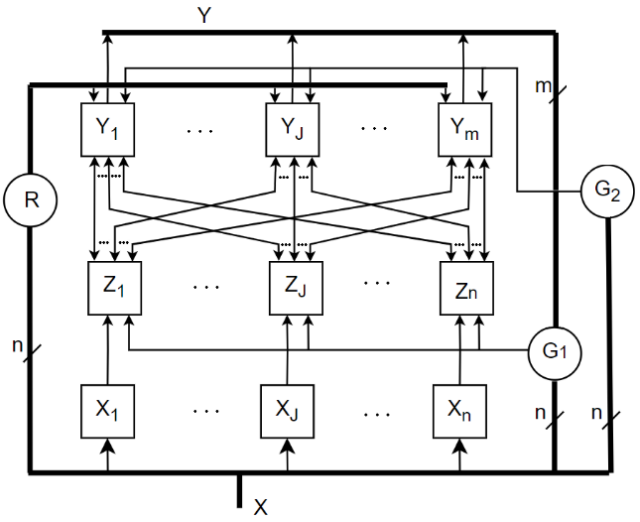

Рис. 1. Архитектура сети АРТ-1

Нейроны Y слоя распознавания получают сигналы от интерфейсных элементов X, нейронов R и G_2. Сигнал G_2 определяется следующим образом: $G_2 = (|X)$. Для активизации нейронов Y и Z необходимо, чтобы как минимум два входных сигнала были активны. В исходном состоянии блоки G_1, G_2 и R находятся в неактивном состоянии. При подаче на вход X входного бинарного вектора некоторые из них перейдут в активное состояние и переведут в возбуждение соответствующие нейроны Z. Активизация нейронов Z возбуждает соответствующие нейроны Y.

В слое Y будет определен нейрон с максимальным значением функции соответствия, этот нейрон перейдет в активное состояние и затормозит работу по распознаванию через элемент управления G_1. В результате обнуления сигнала G_1 активными останутся только те нейроны слоя Z, на входах которых установлено единичное значение от интерфейсных элементов X и от победившего нейрона.

Аппаратная реализация сети АРТ-1 выполнялась на базе микросхемы компании Intel FPGA Cyclone 10LP. Сеть описана как параметризованный модуль, разрядность входного и выходного векторов определяется на этапе компиляции. Нейроны Z и Y описаны как отдельные модули, инстанцированные в сеть с использованием оператора генерации. Операции умножения ориентированы на применение встроенных умножителей. RTL вид Y нейрона показан на рис. 2.

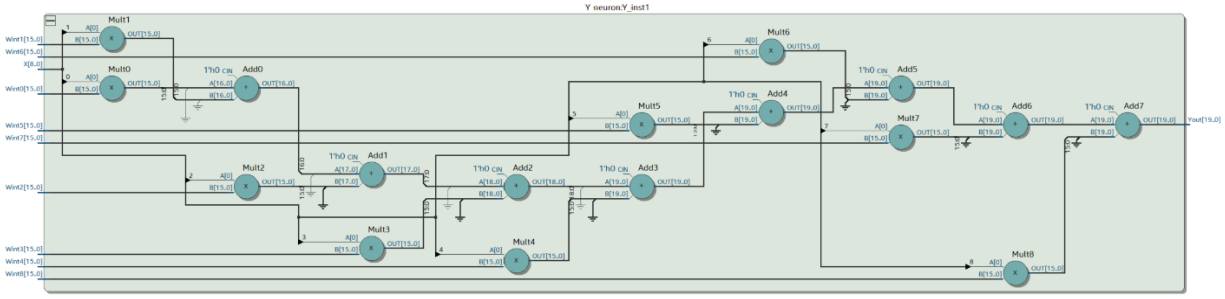

Рис. 2. RTL Y нейрона

При недостатке встроенных блоков умножения, который может возникнуть при увеличении размерности входных и выходных векторов, можно использовать специализированные быстродействующие умножители, реализованные на логических ячейках ПЛИС [14].

Y нейрон реализован в комбинационной форме, триггеры для хранения результатов промежуточных вычислений не предусмотрены, но могут быть введены для повышения производительности сети.

Для определения максимального значения и получения результирующих сигналов Y нейронов был разработан отдельный модуль. На его входы поступают значения функций похожести. Выходной сигнал – вектор Y: значение элемента вектора, индекс которого совпадает с индексом максимального значения устанавливается в 1, а остальные элементы вектора Y обнуляются. Этот сигнал и является результатом работы слоя Y нейронов. Фрагмент RTL модуля сравнения приведен на рис. 3. Пунктирными линиями показаны элементы дерева сравнений.

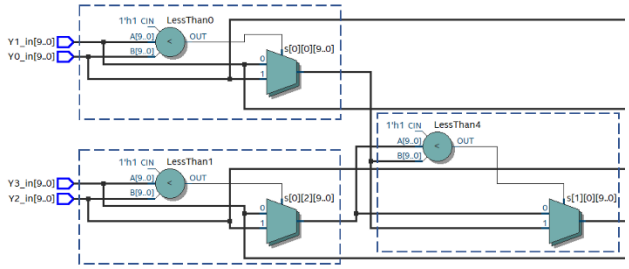

Рис. 3. RTL модуля поиска максимального значения

Для хранения матриц весов сети использовалась распределенная память кристалла FPGA, так как она имеет максимальное быстродействие и допускает одновременную запись большого количества ячеек, что обеспечит максимальное распараллеливание вычислительных процессов.

Рассмотренная комбинационная реализация может быть конвейеризована в части выполнения операций многократного суммирования и поиска максимального значения. Это позволит повысить производительность потоковой обработки входных векторов сетью в целом, потеряв некоторое время на обработку первого входного вектора.

Работа сети была проверена в режиме моделирования с использованием системы ModelSim Altera. Для формирования тестовых сигналов подготовлен набор цифр, представленных в сетке 5 x 9, как показывает рис. 4. Разрядность входного вектора – 45, выходного – 10, разрядность внутренних элементов – 16 бит.

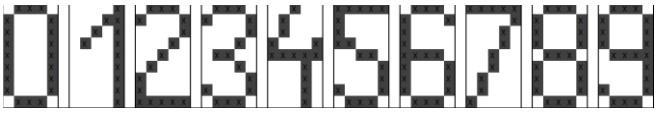

Рис. 4. Фрагмент тестового набора для моделирования работы сети

Моделирование работы устройства показало, что классификатор, обученный на базовом наборе при установленном пороге R_t = 0,875 (двоичный код 0,111) безошибочно определяет цифры, совпадающие с образцами. Снижение порога до R_t = 0,625 (двоичный код 0,101) приводило к смешиванию образов цифр 6,8,9. Рассмотренный пример распознавания изображений подтверждает эффективность использования спроектированной аппаратной реализации АРТ-1 сети.

IV. Заключение

Предложенная аппаратная реализация сети работоспособна и обеспечивает эффективную классификация. По сравнению с программными решениями, аппаратная реализация легко распараллеливается, что позволяет получить аппаратные модули, отличающиеся высоким быстродействием.

Разработанное описание сети на языке SystemVerilog обеспечивает легкое масштабирование за счет использования параметров, характеризующих разрядности данных и количество нейронов в рабочих слоях. При этом увеличение разрядности и данных и слоев практически не скажется на быстродействии сети, так как приведет лишь к увеличению количества параллельно работающих модулей. При этом масштабирование программной реализации приводит к снижению скорости работы.

Использование разработанной сети на базе ПЛИС позволит создавать быстродействующие системы распознавания, пригодные для размещения на периферийных аппаратных узлах и работающие в режиме реального времени.

Дальнейшим развитием работы является исследование процессов масштабирования сети для определения точных зависимостей временных характеристик от размеров сети, а также перенос предложенных цифровых решений на аналоговую элементную базу.

Список литературы

[1] Grossberg S. Adaptive pattern classification and universal recoding. II. Feedback, expectation, olfaction, and illusions // Biological Cybernetics. 1976. Т. 23(4). С. 187–202.

[2] Grossberg S. Conscious Mind, Resonant Brain: How each brain makes a mind. Оксфорд: Oxford University Press, 2021.

[3] Tripathi A., Srivastava G., Singh K. K., Maurya P. K. Review of Unsupervised Adaptive Resonance Theory // 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2019, pp. 409–415. DOI: 10.1109/CONFLUENCE.2019.8776941

[4] Carpenter G. A., Grossberg S. A massively parallel architecture for a self-organizing neural pattern recognition machine // Computer Vision, Graphics, and Image Processing. 1987. Т. 37, вып. 1. С. 54–115.

[5] Zhong C., Songxiang X., Lin L. Simulation of the Concepts of Assimilation, Accommodation and Equilibration in Schema Theory Based on ART2 Network // 7th International Conference on Intelligent Human-Machine Systems and Cybernetics, Hangzhou, China, 2015, pp. 16-20. DOI: 10.1109/IHMSC.2015.274

[6] Carpenter G. A., Grossberg S., Rosen D. B. Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system // Neural Networks. 1991. Т. 4, вып. 6. С. 759–771. DOI: 10.1016/0893-6080(91)90056-B

[7] Bernardes H., Tonelli-Neto M., Minussi C.R. Fault Classification in Power Distribution Systems Using Multiresolution Analysis and a Fuzzy-ARTMAP Neural NetworkAnalysis and a Fuzzy-ARTMAP Neural Network // IEEE Latin America Transactions. 2021. Т. 19, вып. 11. С. 1824–1831. DOI: 10.1109/TLA.2021.9475615

[8] Anagnostopoulos G.C., Georgiopulos M. Hypersphere ART and ARTMAP for unsupervised and supervised, incremental learning // Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium, Como, Italy, 2000, Т.6. С. 59–64. DOI: 10.1109/IJCNN.2000.859373

[9] Tscherepanow M. TopoART: A Topology Learning Hierarchical ART Network // Artificial Neural Networks – ICANN 2010. ICANN 2010. Lecture Notes in Computer Science. Vol 6354. Springer, Berlin, Heidelberg. DOI:10.1007/978-3-642-15825-4_21

[10] Guzhva A., Dolenko S., Persiantsev I. Multifold Acceleration of Neural Network Computations Using GPU // Artificial Neural Networks – ICANN 2009. ICANN 2009. Lecture Notes in Computer Science. Vol 5768. Springer, Berlin, Heidelberg. DOI:10.1007/978-3-642-04274-4_39

[11] Han J., Li Z., Zheng W., Zhang Y. Hardware implementation of spiking neural networks on FPGA // Tsinghua Science and Technology. 2020. Vol. 25, no. 4, pp. 479–486. DOI: 10.26599/TST.2019.9010019

[12] Venkataramanaiah S.K., Yin S., Cao Y., Seo J.-S. Deep Neural Network Training Accelerator Designs in ASIC and FPGA // 2020 International SoC Design Conference (ISOCC), Yeosu, Korea (South). С. 21–22. DOI: 10.1109/ISOCC50952.2020.9333063

[13] Purnomo D.M.J., Arinaldi A., Priyantini D.T., Wibisono A., Febrian A. Implementation of serial and parallel bubble sort on FPGA // Jurnal Ilmu Komputer Dan Informasi. 2016. Т. 9(2). С. 113–120. DOI:10.21609/jiki.v9i2.378

[14] Миронов С.Э., Буренева О.И., Зибарев К.М. Быстродействующие умножители для аппаратной реализации искусственных нейронных сетей // Проблемы разработки перспективных микро- и наноэлектронных систем (МЭС). 2022. Вып. 4. С. 109-116. DOI:10.31114/2078-7707-2022-4-109-116