ВЫБОР АППАРАТНОЙ ПЛАТФОРМЫ ДЛЯ РЕАЛИЗАЦИИ АЛГОРИТМОВ НА ОСНОВЕ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ

© 2016 г. Т.А. ДЕМЕНКОВА, Е.В. ШПИЕВА

Московский технологический университет (МИРЭА)
e-mail: demenkova@mirea.ru, elizaveta.shpieva@smail.inf.h-brs.de

Введение

Искусственные нейронные сети (ИНС) в настоящее время находят широкое применение в самых разных предметных областях. Они используются для идентификации статических и динамических объектов, для построения ассоциативной памяти и моделей временных рядов, сжатия информации, в системах поддержки принятия решений, для целей прогнозирования, классификации и распознавания образов, а также как средство решения некоторых задач вычислительного характера. ИНС, построенные по принципу организации и функционирования биологических нейронных сетей по своей природе требуют сложных параллельных вычислений. Таким образом, высокая скорость работы в приложениях реального времени и обучение больших наборов данных могут быть достигнуты только в том случае, если сети реализованы с использованием параллельной аппаратной архитектуры [1].

Преимущества ИНС для реализации на аппаратных средствах

Искусственная нейронная сеть - это математическая модель, основанная на биологических нейронных сетях. Существуют различные типы искусственных нейронных сетей, такие как упреждающие нейронные сети, многослойные и рекуррентные нейронные сети. Каждая из них имеет свои недостатки и достоинства и может быть применена в различных областях: распознавание лиц, предсказание последовательностей, процессы управления, фильтрации и адаптации.

Биологическая нейронная сеть очень развита и сложна. Для построения математической модели ИНС необходимо сделать несколько допущений. Каждый нейрон обладает некоторой передаточной функцией, определяющей условия его возбуждения. При прохождении синапса сигнал меняется линейно, то есть сила сигнала умножается на некоторое число. Это число будем называть весом синапса или весом соответствующего входа нейрона. Деятельность нейронов синхронизирована, то есть время прохождения сигнала от нейрона к нейрону фиксировано и одинаково для всех связей.

Рассмотрим основные типы архитектуры нейронных сетей: многослойные персептроны (multilayer percetron) и рекуррентные сети (РНС). В многослойных персептронах множество сенсорных элементов составляют входной слой нейронной сети (входных узлов или узлов источника). Имеются также один или нескольких скрытых слоев вычислительных нейронов (hidden layer) и один выходной слой нейронов (output layer). Входной сигнал распространяется по сети в прямом направлении, от слоя к слою. Такие сети широко используются в распознавании образов. В рекуррентных нейронных сетях нейроны скрытого слоя могут иметь сигналы, передвигающиеся в обоих направлениях с помощью введения петель в сети. На каждом такте работы нейронной сети один или несколько нейронных элементов обрабатывают информацию, полученную на предыдущем шаге, то есть функционирование рекуррентных НС носит итеративный

характер [2]. Такая обработка информации происходит до тех пор, пока все нейронные элементы не перейдут в состояние равновесия. При этом состояния нейронных элементов перестают изменяться. С помощью РНС можно решать нелинейные динамические задачи [3]. Скрытый слой, называемый резервуаром сети данного типа, представляется "черным ящиком", для которого могут быть заданы различные параметры, дающие общее представление о резервуаре нашей сети. Возможна установка небольшого количества параметров, таких как размер резервуара, разреженность входных соединений и хранилища, диапазон значений весов для входов и резервуара, а также масштабирование резервуара матрицы [4].

Нейронная сеть является надежным инструментом для адаптации к изменениям в среде благодаря своей возможности справляться с динамическими изменениями [5]. Она может использоваться, например, в нейросетевом контроллере для небольшого автономного робота. Другие адаптивные системы, такие, как итерационный контроллер, имеют свои недостатки: низкую сходимость к решению и также необходимость иметь большой объём данных для обучения сети. Нейронные сети сходятся к решению быстрее и при этом могут работать в нелинейных системах [6].

Можно выделить следующие преимущества рекуррентных нейронных сетей:

работа с динамическими изменениями;

быстрая сходимость;

более точное отображение входа на выход.

Если говорить об аппаратной реализации такой сети, то можно выделить следующие положительные характеристики [7]:

параллельная обработка, каждый нейрон обрабатывается на отдельном блоке;

асинхронные операции, нейроны ИНС могут работать на максимальной скорости оборудования;

редакция ошибок, если некоторые из нейронов будут выдавать ошибочную информацию, они могут быть удалены без каких-либо проблем;

регулярность, нейроны строятся из простых блоков, таких, как сложение и сдвиг;

высокая скорость обработки данных.

Далее рассмотрим различные аппаратные носители для реализации наилучшей модели нейронной сети, которая потенциально могла бы строить лучший прогноз поведения и быстро адаптироваться к внешним изменениям.

Аппаратные реализации ИНС обычно называют аппаратными нейронными сетями (АНС). Неформально АНС можно рассматривать как устройства, предназначенные для реализации нейронных архитектур и алгоритмов обучения с преимуществами, связанными с параллельной природой ИНС. Для покрытия большинства реальных проблем от нейронных сетей требуется как минимум от 100 до 1000 узлов, а число межсоединений обычно колеблется в пределах от 104 до 106.

ИНС, как правило, определяется в терминах топологии сети, алгоритмом обучения, количеством и типом входов/выходов, количеством процессорных элементов (ПЭ) и синаптических соединений, количеством слоев и т.д. Для аппаратной реализации, кроме этого, могут существовать дополнительные технические характеристики в используемой технологии (аналоговой, цифровой, или гибридной): представление данных (с фиксированной или плавающей запятой), точность, программируемые или фиксированные соединения, обучение на чипе или чип-в-цикле, передаточная функция на кристалле или вне кристалла, например, таблицы поиска (LUT), и количество каскадов.

Большинство доступных микросхем АНС являются цифровыми и большинство из них используют КМОП-технологии. Есть несколько категорий цифровых микросхем с различной архитектурой, такие как многокристальный секционный микропроцессор (MCM), MISD архитектура параллельных вычислений и систолические матрицы (СМ). Преимущества цифровой технологии включают в себя использование хорошо известных методов изготовления и гибкий дизайн.

Самой большой проблемой для проектировщиков является реализация синапса, который обычно является самым медленным элементом. У каждой категории есть

свои достоинства и недостатки.

MCM архитектура обычно включает в себя обучения вне кристалла. В случае SIMD (одиночный поток команд) каждый из процессорных элементов обрабатывает одни и те же инструкции одновременно, но на разных наборах данных.

При реализации на кристалле с архитектурой MISD каждый из процессорных элементов делает один шаг вычисления синхронно с другими процессорными элементами и затем передает свой результат следующему процессору, что делает архитектуру подходящей для реализации матричного умножения, которое является общим для построения ИНС.

Примеры реализации

Исследования конкретных реализаций нейронных сетей позволяют сделать вывод, что в каждом решении на первое место выступает проблема возможности выполнения данного алгоритма на данной платформе. Например, в работе [8], представлена аппаратная модель для морфологической нейронной сети, которая заменяет классические операции умножения и сложения на операции сложения и операции получения максимума или минимума. Известна модель синапса с использованием КМОП-структуры в стандартном процессоре, выполненном по технологии 0,35-мкм. Конструкция синапса затем используется в архитектуре СБИС, состоящей из 2176 синапсов для приложения выделения признаков отпечатка пальца.

В работе [9] описывается реализация схемы обработки сигналов для непрерывной рекуррентной нейронной сети с использованием аналоговой СБИС в смешанном режиме, где переменные состояния представлены напряжениями, в то время как нервные сигналы передаются в виде токов. Использование тока увеличивает точность нейросигналов и позволяет поддерживать большие расстояния между нейронами, что делает эту архитектуру относительно надежной и масштабируемой.

Первая модель АНС на основе запоминающего устройства с произвольным доступом (ЗУПВ) состоит из ПЭ (нейронов), которые имеют только дискретные входы и выходы и никакого веса между узлами. Нейронные функции сохраняются в таблицах поиска (LUT). В отличие от других моделей нейронных сетей они могут обучаться очень быстро и могут быть реализованы на простом оборудовании. Вместо того, чтобы регулировать веса в обычном смысле, АНС на основе ЗУПВ обучаются путем изменения содержимого таблиц поиска. АНС на основе ЗУПВ нашли широкое применений, включая построение систем распознавания образов [10].

Реконфигурируемые программируемые логические интегральные схемы (ПЛИС) являются эффективным программируемым ресурсом для реализации АНС, позволяющим различные варианты проектирования, которые можно оценивать в течение очень короткого времени. Они имеют низкую стоимость, легко доступны, а их свойства реконфигурации предлагают гибкость программного обеспечения.

Важной проблемой, стоящей перед разработчиками АНС на основе ПЛИС, является выбор подходящей модели ИНС для конкретной микросхемы, которую необходимо реализовать с использованием оптимальных аппаратных ресурсов. Был проведен сравнительный анализ требований к оборудованию для реализации четырех моделей ИНС на программируемых логических схемах с архитектурой FPGA. Выбранные модели включали в себя многослойный персептрон MLP с алгоритмом BP (Back Propagation) и RBF (Radial Basis Function) сети классических моделей, а также две модели SNN - LIF и Spike Response. Эти модели были затем проанализированы на эталонной задаче классификации аппаратных ресурсов ПЛИС. Результаты исследования свидетельствуют о том, что LIF модель могла бы быть наиболее подходящим выбором реализации для нелинейных задач [11].

Заключение

В результате выполненного в работе анализа возможности выбора аппаратной платформы для реализации искусственных нейронных сетей можно сделать следую-

щие выводы. Несмотря на существование языков описания аппаратуры высокого уровня, как например, Verilog и VHDL, и компиляторов, для аппаратных проектов нейронных сетей все еще требуются усилия и специальные знания со стороны разработчиков, чтобы оптимально использовать все имеющиеся ресурсы для достижения высокой скорости и низкой рассеиваемой мощности. Правильное отображение моделей искусственных нейронных сетей на параллельных архитектурах, которые обеспечивают эффективное вычисление и коммуникации, таким образом является ключевым шагом в создании любой конструкции АНС. Однако существует необходимость в разработке инструментов для автоматического перевода моделей высокого уровня ИНС на аппаратные средства.

В настоящее время довольно привлекательным вариантом элементной базы для рассмотренных задач является аппаратная реализация на программируемых логических интегральных схемах с архитектурой FPGA, которая позволяет воспользоваться такими преимуществами ПЛИС как аппаратная эффективность и гибкость программного обеспечения. Кристаллы ПЛИС представляет собой очень мощный вариант для реализации аппаратных нейронных сетей, поскольку они действительно могут использовать свои возможности параллельной обработки информации, тем самым увеличивая скорость обработки всей системы.

Исследования проводились в рамках выполнения проекта №35 базовой части государственного задания №2014/112 Министерства образования и науки.

СПИСОК ЛИТЕРАТУРЫ

1. *Sahin S, Becerikli Y. and Yazici S.* Neural network implementation in hardware using FPGAs. // Proceedings of the 13th international conference on Neural information processing - Volume Part III (ICONIP'06), (2006) P. 1105-1112.
2. *Хайкин С.* Нейронные сети: полный курс.Пер. с англ. – М.: Изд. дом "Вильямс", 2006.
3. *Schrauwen B., Verstraeten D. and Campenhout J. Van.* An overview of reservoir computing: theory, applications and implementations // Proceedings of the 15th European Symposium on Artificial Neural Networks. 2007. P. 471-482.
4. *Rodan A. and Tino P.* Minimum complexity echo state network. // IEEE transactions on neural networks a publication of the IEEE Neural Networks Council, т. 22, № 1, 2011. PP. 131-144.
5. *Atiya A.F. and Parlos A.* New results on recurrent network training: unifying the algorithms and accelerating convergence // IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council, т. 11, № 3, 2000, P. 697-709.
6. *Rovithakis G.A. and Christodoulou M. A*. Adaptive Control with Recurrent High-order Neural Networks: Theory and Industrial Applications. // Advances in Industrial Control. Springer-Verlag. 2000 P. 5-10.
7. *Misra J. and Saha I.* Artificial neural networks in hardware: A survey of two decades of progress // Neurocomput.74, 1-3 (December 2010), P. 239-255.
8. *Ortiz J., Ocasio C.* Analog hardware model for morphological neural networks // Proceeding of International Conference on Neural Networks and Computational Intelligence, 2003.
9. *Brown B., Yu X., Garverick S.* Mixed-mode analog VLSI continuous-time recurrent neural network // Proceedings of International Conference on Circuits, Signals and Systems, 2004.
10. *Kim D., Kim H., Han G., Chung D.* SIMD neural network processor for image processing, // Advances in Neural Networks ISNN 2005 3497/2005 (2005) 665-672.
11. *Johnston S., Prasad G., Maguire L.P., McGinnity T.M.* Comparative investigation into classical and spiking neuron implementations on FPGA // ICANN (1), 2005, pp. 269-274.