УДК 00.004.27 + 004.3

ПРОБЛЕМЫ РЕАЛИЗАЦИИ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ НА FPGA

Гонтаренко Б.В.
Донецкий национальный технический университет

Рассматриваются вопросы реализации нейронных сетей в целом и основной их составляющей – искусственного нейрона – в частности. Описываются различия аппаратной и программной реализации нейронных сетей, а так же обосновывается удобство реализации нейронных сетей на FPGA. Приводится информация о проблемах, возникающих при реализации нейронных сетей и об их возможных решениях.

Введение

Искусственная нейронная сеть (ИНС) – математическая модель, а также её программная или аппаратная реализация, построенная по принципу организации и функционирования биологической нейронной сети – сети нервных клеток живого организма. ИНС представляют собой систему соединённых и взаимодействующих между собой простых процессоров (искусственных нейронов). Такие процессоры обычно довольно просты, особенно в сравнении с процессорами, используемыми в персональных компьютерах [1].

Целью доклада является рассмотрение основных сложностей и проблем, с которыми может столкнуться разработчик искусственной нейронной сети с реализацией её на FPGA (ПЛИС, программируемая логическая интегральная схема), а также возможных путей их решения.

Актуальность поставленных задач подтверждается тем, что развитие концепций реализации искусственных нейронных сетей на базе программируемых логических схем (в большей степени из-за удобства работы с ними и возможности быстрого их программирования) только набирает обороты.

Далее будут рассмотрены, различия в программной и аппаратной реализации искусственных нейронных сетей и проблемы, возникающие при аппаратной реализации на FPGA.

1 Программная и аппаратная реализации ИНС

Существуют два способа реализации искусственных нейронных сетей – аппаратная и программная. Программная реализация, уступая аппаратной по скорости работы и автономности, обладает рядом очевидных преимуществ, связанных с простотой использования и внедрения в информационно-управляющую систему. Это вполне логично: нейронные сети – это относительно новая область, а создание ПО — достаточно гибкий процесс, что позволяет тестировать и внедрять с малыми затратами некоторые экспериментальные методы.

Однако специализированные аппаратные средства предлагают заметные преимущества в определенных ситуациях (чаще всего – в плане скорости). К основным достоинствам аппаратной реализации ИНС перед программным исполнением можно отнести скорость (увеличивается за счет аппаратной реализации параллельных вычислений), надежность (вероятность отказа аппаратуры меньше вероятности сбоя программы), безопасность (в плане защиты авторских прав) и дополнительные режимы эксплуатации.

Что касается аппаратной реализации нейронных сетей на FPGA, то она выгодно отличается от реализации на специальных DSP-процессорах (поскольку они выпускаются серийно) и от реализации на ASIC-микросхемах (поскольку они не подлежат переконфигурированию). Реализация на FPGA наиболее точно передает параллельную архитектуру нейронов и предоставляет возможность гибкого реконфигурирования всей нейронной сети и её составляющих – искусственных нейронов [2]. Так же FPGA – это сравнительно доступные схемы небольшой стоимости, что позволяет быстро и недорого реализовать всю систему. Кроме того, конфигурацию основанных на FPGA нейронных сетей легко изменить.

1

2 Проблема реализации ИНС, связанная с аппаратными затратами

Модели искусственных нейронных сетей, как правило, во многом зависят от массивных параллельных вычислений. Поэтому, чтобы обеспечить высокую скорость работы в режиме реального времени, нейронные сети должны быть реализованы с помощью параллельных аппаратных архитектур.

Для аппаратной реализации в целом очень важно отделять фазы обучения и восстановления нейронной сети. В общем, архитектура ИНС состоит из набора входов и взаимосвязанных нейронов специальной структуры. Нейрон можно считать базовым элементом обработки информации, и его структура определяет сложность сети в целом.

Фундаментальная проблема ограничения размеров нейронной сети на FPGA заключается в стоимости реализации умножений, связанных с синоптическими связями – для полностью параллельной ИНС количество умножителей должно быть равно количеству нейронов в сети.

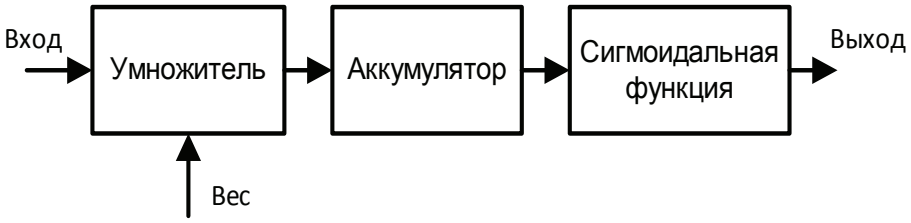

Рисунок 1. Типичная архитектура нейрона

На FPGA с интеграцией в 40 тысяч эквивалентных вентилей можно реализовать до 15 параллельно работающих нейронов, но доступны также FPGA с интеграцией в 2 миллиона вентилей, что дает возможность реализовать, соответственно, 1000 параллельных нейронов. Такого числа нейронов уже достаточно для разработки серьезных нейроприложений. Но при необходимости реализации сложных нейронов ситуация сильно ухудшается.

Хотя прототип и может быть создан на базе FPGA, предлагающего большое количество умножителей, общая цель заключается в использовании как можно меньшего количества ресурсов. Практическая реализация ИНС осуществляется либо за счет сокращения числа умножителей, либо их упрощением.

Один из способов уменьшения количества умножителей – это обеспечение возможности использования всеми входами нейрона одного и того же умножителя [3]. Другой способ уменьшения количества элементов схемы, необходимых для реализации операции умножения – это базирующийся на последовательности битов метод стохастических вычислений.

Что касается сумматоров, то их сложность зависит от точности входов синапсов и общего количества входов нейрона.

3 Проблема реализации ИНС, связанная с внутренним представлением функций активации нейрона

Еще одной проблемной областью, имеющей особое значение для аппаратной реализации ИНС, является система передаточных функций активации нейрона. Передаточная функция определяет зависимость сигнала на выходе нейрона от взвешенной суммы сигналов на его входах. В большинстве случаев она является монотонно возрастающей и имеет область значений (−1,1) или (0,1), однако существуют исключения. Также для некоторых алгоритмов обучения сети необходимо, чтобы она была непрерывно дифференцируемой на всей числовой оси. Искусственный нейрон полностью характеризуется своей передаточной функцией. Использование различных передаточных функций позволяет вносить нелинейность в работу нейрона и в целом нейронной сети [4].

Сигмоидальная функция, традиционно используемая в ИНС, не подходит для прямой цифровой реализации, поскольку состоит из бесконечного ряда экспонент. Поэтому, в большинстве

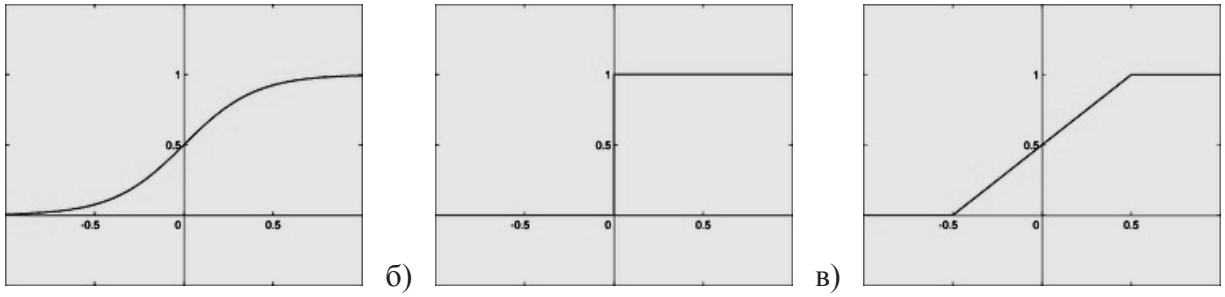

Рисунок 2. Функции активации нейрона: а) сигмоидальная; б) пороговая; в) линейная

реализаций разработчики прибегают к различным методам аппроксимации сигмоидальной функции с помощью аппаратных поисковых таблиц (lookup table, LUT). В них, собственно, и хранятся примеры сигмоидальных функций, нужных для аппроксимации. Тем не менее, количество требуемых для реализации подобного подхода аппаратных средств может быть чрезмерно велико, особенно если требуется очень точная аппроксимация.

В других вариантах реализации цифрового приближения сигмоидальной функции используются сумматоры, регистры сдвига и множители.

4 Проблема выбора точности веса функции активации

Выбор точности веса – это одно из самых важных решений при реализации искусственных нейронных сетей на FPGA. Выбор точности веса используется для выбора нужного соотношения между возможностями реализованной нейронной сети и стоимостью её реализации. Более высокая точность веса означает меньшее количество ошибок квантования в финальной реализации, а низкая точность приводит к более простой структуре, большей скорости работы и уменьшению потребляемой энергии и занимаемого пространства на кристалле.

Один из способов достижения компромисса заключается в определении минимальной точности, необходимой для решения заданной проблемы. Традиционно, минимальная точность находится методом «проб и ошибок», при помощи моделирования нейронной сети в программной среде до её фактической аппаратной реализации.

Холт и Бейкер [5] в своей работе исследовали минимальную точность, требуемую для класса эталонных классификационных проблем, и пришли к выводу, что 16-разрядная переменная с фиксированной точкой и является той самой минимальной допустимой точностью, которая практически не влияет на способность ИНС к обучению.

В последнее время более распространенным стал теоретический подход к выбору точности веса. В этом случае нужно понимать, что «трудность» решения данной задачи (т.е. насколько трудно она решается) необходимо связывать с требуемым количеством весов и их необходимой точностью.

Выводы

При реализации ИНС на FPGA следует понимать, какую роль играет возможность реконфигурации аппаратных средств и разрабатывать стратегии для эффективного использования аппаратных ресурсов программируемых интегральных схем. Немаловажными и проблемными этапами разработки нейронной сети являются: реализация нейрона в целом и привязанного к нему умножителя в частности; реализация функции активации нейрона и её аппроксимация; выбор точности весовых коэффициентов входных сигналов.

Литература

[1] Искусственная нейронная сеть. Материал из Википедии – свободной энциклопедии. Электронный ресурс. Режим доступа: http://ru.wikipedia.org/wiki/Искусственная_нейронная_сеть

[2] Сунд Су Ким и Сеул Джунг, «Аппаратная реализация контроллера нейронной сети реального времени на DSP и FPGA», доклад на международной конференции IEEE.

[3] С.Л. Блейд и Б.Л.Хатчингс, «Реализация стохастической нейронной сети на FPGA», доклад на секции IEEE на конференции FPGAs for Custom Computing Machines.

[4] Искусственный нейрон. Материал из Википедии – свободной энциклопедии. Электронный ресурс. Режим доступа: http://ru.wikipedia.org/wiki/Искусственный_нейрон

[5] Холт Дж.Л., Бейкер Т.Е. «Моделирование обратного распространения с использованием вычислений с ограниченной точностью», Международная конференция по нейронным сетям, 1991.