1.4. ПРИМЕНЕНИЕ ВЫЧИСЛИТЕЛЬНЫХ СРЕД ДЛЯ УСКОРЕНИЯ РЕКУРРЕНТНЫХ НЕЙРОННЫХ СЕТЕЙ

Шатравин В. Шашев Д. В.
Томский государственный университет (ТГУ), Томск, Россия

Одним из основных факторов, ограничивающих применение современных алгоритмов машинного обучения в технических системах, является несовершенство используемого аппаратного обеспечения. Особенно остро проблема стоит для крупных нейронных сетей в маломощных и автономных системах, имеющих жесткие ограничения к массе и энергопотреблению. Большинство предлагаемых на сегодняшний день аппаратных ускорителей нейронных сетей либо имеют высокое энергопотребление и массу, либо поддерживают лишь очень ограниченное множество алгоритмов. Решением этой проблемы может быть применение перестраиваемых аппаратных ускорителей, которые поддерживают динамическую настройку на реализацию требуемых алгоритмов. Одним из способов построения таких ускорителей могут быть решения на основе концепции перестраиваемых вычислительных сред (ПВС). В данной работе представлена реализация рекуррентных архитектур нейронных сетей на примере сети Хопфилда и сети долгой краткосрочной памяти (LSTM) на ускорителях, построенных на основе ПВС. Приведены формулы оценки быстродействия разработанных моделей на основе результатов симуляций на FPGA. Полученные оценки показывают высокое быстродействие предложенных моделей в сравнении с существующими аналогами при значительно большей занимаемой на полупроводнике площади. Согласно оценкам, расчёт одного шага LSTM сети с 25 скрытыми нейронами занимает 223 нс. Результаты позволяют сделать вывод о большом потенциале применения перестраиваемых сред для реализации рекуррентных сетей и необходимость дальнейших оптимизаций предложенных моделей.

Введение

Активное развитие алгоритмов машинного обучения открывает перед техническими системами новые возможности в области решения задач распознавания образов, обработки естественного языка, предсказательной аналитики и многих других. В первую очередь это достигается при помощи глубоких свёрточных и рекуррентных нейронных сетей. В рамках данной работы наибольший интерес представляют рекуррентные сети, отличительной особенностью которых является наличие обратных связей между нейронами. Обратные связи создают эффект памяти, что позволяет рекуррентным сетям накапливать информацию, то есть учитывать результаты прошлых итераций при обработке новых входных сигналов. Такие сети наиболее эффективны при работе с данными, представленными в виде последовательности – устная речь, текст, динамика котировок, результаты измерений датчика. На сегодняшний день рекуррентные сети представляют собой широкий класс разнообразных архитектур, среди которых наиболее известной являются сети с долгой краткосрочной памятью (LSTM).

В то же время хорошо известна проблема вычислительной сложности алгоритмов машинного обучения. Современные нейронные сети могут насчитывать сотни миллиардов параметров, что делает применение классических вычислительных устройств на основе центральных процессоров неэффективным с точки зрения скорости вычислений и затрачиваемой энергии.

В связи с этим на практике большую популярность получили специализированные под нейросетевые алгоритмы вычислительные устройства, архитектура которых учитывает особенности осуществляемых преобразований. Такие вычислительные устройства называют аппаратными ускорителями нейронных сетей. Аппаратные ускорители показывают высокую эффективность в алгоритмах машинного обучения благодаря распараллеливанию матричных операций и оптимизации доступа к памяти. В качестве аппаратных платформ таких устройств выступают графические процессоры (GPU), программируемые логические интегральные схемы (FPGA) и интегральные схемы специального назначения (ASIC) [Chen; 2019; He, 2021; Ghimire, 2022]. К последним, в частности, относятся тензорные процессоры компаний Google и Huawei, ускоритель Eyeriss, а также многие другие. Каждая из перечисленных платформ имеет свои особенности и предпочтительные области применения.

В рамках данной статьи будут рассматриваться аппаратные ускорители для широкого класса маломощных мобильных и автономных систем, к которым относятся БПЛА, смартфоны, мобильные роботы, спутниковые системы, умные датчики интернета вещей и многие другие. Для таких систем наиболее характерно применение вычислителей на основе FPGA и ASIC в связи с их высокой производительностью при умеренном энергопотреблении. Однако анализ показывает, что большинство существующих решений специализируются на узком классе архитектур нейронных сетей и поддерживают относительно небольшой набор алгоритмов. В то же время сложные мобильные системы могут нуждаться сразу в нескольких нейронных сетях с различными архитектурами, что приводит к необходимо-

сти использования сразу нескольких ускорителей, каждый из которых будет специализирован под свой класс нейронных сетей.

Одним из путей решения этой проблемы может быть применение перестраиваемых аппаратных ускорителей, которые поддерживают широкий класс алгоритмов, но в каждый момент времени реализуют только некоторые из них. В основе архитектуры таких ускорителей могут лежать перестраиваемые вычислительные среды (называемые также однородными), которые обеспечивают не только высокую степень параллельности вычислений, но и динамическую низкоуровневую настройку отдельных участков, и высокую надёжность.

Однако применение аппаратных ускорителей на перестраиваемых средах для реализации нейронных сетей конкретных архитектур требует разработки специализированных алгоритмов. В работах [Shatravin, 2021, 2022] предложены алгоритмы для реализации глубоких сетей прямого распространения, а также сигмоидной функции активации нейронов. В данной работе будут представлены алгоритмы реализации на перестраиваемых средах рекуррентных нейронных сетей на примере сети Хопфилда и сети с долгой краткосрочной памятью (LSTM).

1. Рекуррентные нейронные сети

Рекуррентные нейронные сети – широкий класс архитектур нейронных сетей, отличительной особенностью которых является использование полученных на выходе сети результатов при обработке новых входных сигналов. Это достигается благодаря наличию обратных связей, через которые выходной сигнал после некоторых преобразований попадает обратно на вход сети, где проходит обработку вместе с новым входным сигналом. Это позволяет рекуррентным сетям эффективно работать с данными, представляющие собой последовательность – динамика изменения параметров некоторой системы, речь, временные ряды и т.д.

По характеру связей между входными и выходными сигналами выделяют несколько типов рекуррентных сетей: «один вход к одному выходу», «один вход ко многим выходам», «многие входы к одному выходу», «многие входы к многим выходам» (рис. 1).

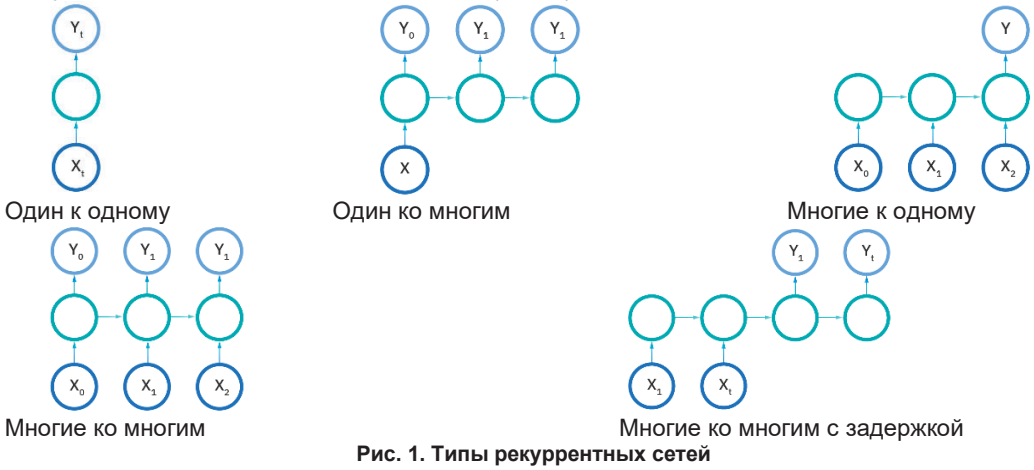

Рис. 1. Типы рекуррентных сетей

В частности, рассматриваемая далее сеть Хопфилда относится к типу «один к одному», так как каждому входному сигналу соответствует одно устойчивое состояние, к которому за некоторое количество шагов сойдётся сеть, а LSTM – к типу «много ко многим», так как на каждом такте на сеть подаётся новый входной сигнал и выводится текущий результат.

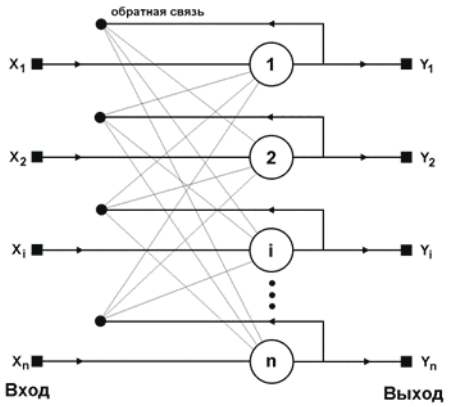

Рис. 2. Рекуррентная сеть Хопфилда

Упомянутая ранее рекуррентная сеть Хопфилда – одна из классических архитектур рекуррентных нейронных сетей. Она состоит из одного слоя нейронов, причём выход каждого нейрона соединён с входами всех остальных нейронов через обратные связи (рис. 2). Сеть Хопфилда имеет симметричную матрицу связей, оперирует двоичными одноразрядными сигналами и использует пороговую функцию активации нейронов (рис. 3). Попав на вход такой сети, сигнал циклически перемещается в ней до перехода сети в некоторое устойчивое состояние, то есть состояние, при котором значение сигнала (энергия сети) прекращает дальнейшие изменения. Полученный результат соответствует одному из образов, на которые данная сеть была обучена. Иными

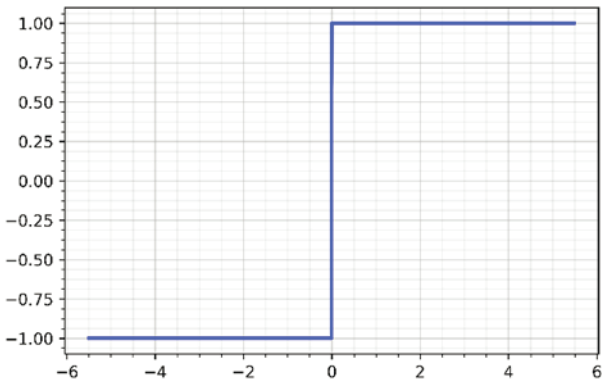

Рис. 3. Пороговая функция активации нейронов в сети Хопфилда

словами, сеть Хопфилда выполняет функцию автоассоциативной памяти – восстанавливает исходный образ до одного из заданных шаблонов. В связи с тем, что по мере обработки сигнала энергия сети уменьшается, сеть Хопфилда находит применение в решении задач оптимизации

Гораздо более широкое распространение получили рекуррентные сети с долгой краткосрочной памятью (Long short-term memory, LSTM). LSTM сети относятся к типу «многие ко многим», которые на каждом такте получают новый элемент входной последовательности и возвращают результат, учитывающий не только текущее значение на входе, но и значения, вычисленные на предыдущих тактах. Для многих рекуррентных архитектур характерна проблема «забывания» – результаты с предыдущих тактов относительно быстро заменяются более новыми данными, что препятствует анализу длинных цепочек связанных событий. LSTM сети получили своё название благодаря способности долговременно хранить ключевую информацию с предыдущих тактов и управлять её изменением [Hochreiter, 1997].

Для понимания основных принципов работы LSTM сети рассмотрим её внутреннее устройство (рис. 4) [Olah, 2015]. Долговременное хранение необходимой для работы сети информации обеспечивается введением состояния ячейки (сигнал C_t). На каждом шаге работы сети этот сигнал модифицируется определённым образом: из него устраняется уже неактуальная информация (пурпурный участок сети) и добавляется новая информация, соответствующая текущему входному сигналу x_t (зелёный участок сети). После этого из текущего входного сигнала x_t и состояния ячейки C_t формируется новый выходной сигнал h_t (оранжевый участок сети). Этот процесс повторяется для каждого нового элемента входной последовательности X. Можно видеть, что описанная архитектура сети состоит из нескольких слоёв нейронов с функциями активации сигмоида и гиперболический тангенс, а также операций поточечного сложения и умножения. LSTM сети активно используются при решении широкого класса задач, в том числе для обработки естественного языка, текстов, хронологически упорядоченных событий. В связи с большой сложностью необходимых в данных задачах сетей остро стоит вопрос их аппаратного ускорения.

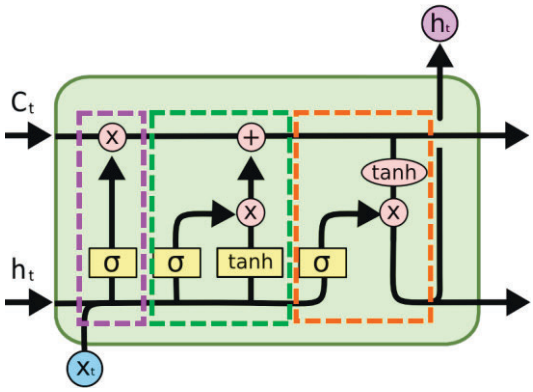

Рис. 4. Строение LSTM сети

2. Аппаратные ускорители на перестраиваемых вычислительных средах

Как было сказано ранее, одним из перспективных направлений развития аппаратных ускорителей нейронных сетей является разработка реконфигурируемых ускорителей на основе перестраиваемых вычислительных сред (ПВС). Благодаря способности к перестраиваемости такие ускорители могут поддерживать широкий класс нейросетевых алгоритмов, динамически настраиваясь на требуемые в данный момент модели сетей. В работах [Shatravin, 2021, 2022] показано, как на ускорителях на основе ПВС могут быть реализованы нейронные сети прямого распространения, а также некоторые основные функции активации. Симуляции разработанных моделей демонстрируют высокое быстродействие. На данный момент эта теория активно исследуется, разрабатываются адаптации других распространённых нейросетевых алгоритмов. Для обсуждения реализации рекуррентных нейронных сетей необходимо предварительно рассмотреть основные аспекты предлагаемых моделей аппаратных ускорителей.

Лежащие в основе ускорителей перестраиваемые вычислительные среды представляют собой модель вычислительного устройства, состоящего из большого

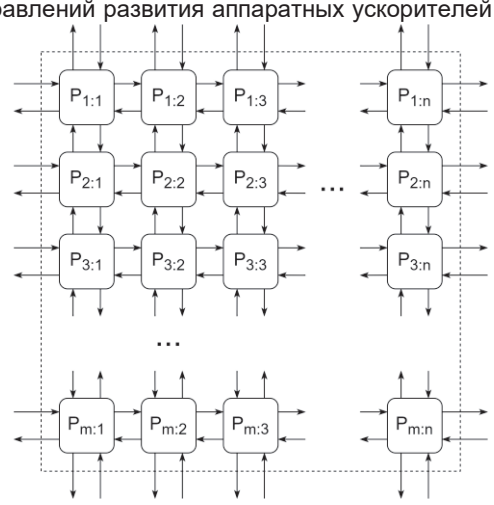

Рис. 5. Перестраиваемая вычислительная среда

количества простых вычислительных элементов (ВЭ). В рамках данной работы будут рассматриваться двумерные среды с квадратными элементами (рис. 5). Все ВЭ одинаковы, но каждый из них может быть независимо от других настроен на выполнение определённой операции из заранее заданного базиса операций. Для хранения текущей настройки (операции и дополнительных параметров) каждый ВЭ имеет небольшой объем внутренней памяти. Соседние элементы соединены друг с другом симметричными связями, что позволяет им обмениваться результатами вычислений и передавать сигнал через себя в другие участки среды.

Применение ПВС к построению аппаратных ускорителей обусловлено их многочисленными преимуществами. Благодаря тому, что каждый элемент среды функционирует независимо от других, среда может выполнять алгоритмы с высокой степенью параллельности, что играет большую роль в алгоритмах, поддерживающих распределённые вычисления и распараллеливание, к которым в том числе относятся и нейросетевые алгоритмы. Способность ПВС к динамической перенастройке позволяет не только реализовать в рамках одного вычислителя широкий класс алгоритмов, но и обеспечивает их изменение в процессе функционирования устройства. Изотропность среды допускает выполнение вычислений на любом её участке, что может быть полезно в случае частичного повреждения устройства, а также допускает одновременное выполнение нескольких различных алгоритмов на разных участках одной среды. Немаловажным преимуществом однородности вычислителей являются упрощение промышленного производства и масштабирования.

Рис. 6. Нейрон из элементов среды (слева) и полносвязный слой из трёх нейронов (справа)

В упомянутых ранее работах [Shatravin, 2021, 2022] предлагается реализация нейрона в виде цепочки из вычислительных элементов ПВС, а полносвязного слоя – как объединения таких цепочек (рис. 6). Этот подход обеспечивает высокую гибкость при реализации требуемых моделей нейронных сетей, так как изменяя длину цепочки элементов можно получить нейрон требуемой конфигурации. В то же время функции активации нейронов могут быть реализованы как в виде отдельной операции базиса элемента (как в случае с очень широко применяемой функцией линейного выпрямителя ReLU), так и в виде определённым образом настроенной группы элементов (сигмоида, гиперболический тангенс и т.д.). Пример реализации сигмоиды на предложенной модели ускорителя показан на рисунке 7 [Shatravin, 2022]. Алгоритм настройки такого ускорителя описан в работе [Шатравин, 2022].

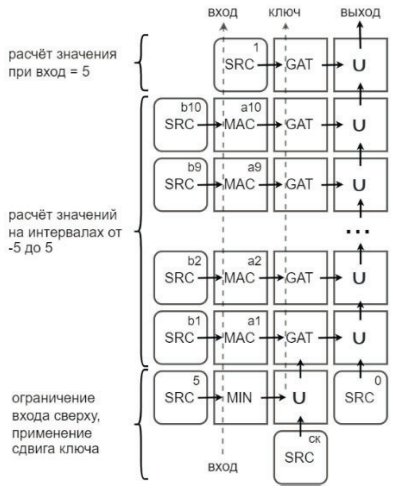

Рис. 7. Реализация сигмоидной функции активации на ПВС

Для достижения лучших результатов может быть использован сегментированный режим функционирования среды. В этом режиме среда разбивается на несколько участков (сегментов), каждый из которых настраивается на реализацию одного из слоёв сети (рис. 8). Входной сигнал последовательно перемещается от сегмента к сегменту, проходя соответствующую обработку, а пройденные сегменты перенастраиваются на последующие слои. Такой подход позволяет уменьшить накладные расходы на перенастройку элементов, улучшить утилизацию ресурсов среды, устранить внешний обмен промежуточными результатами, а также реализовать сеть с произвольным количеством слоёв на среде ограниченного размера.

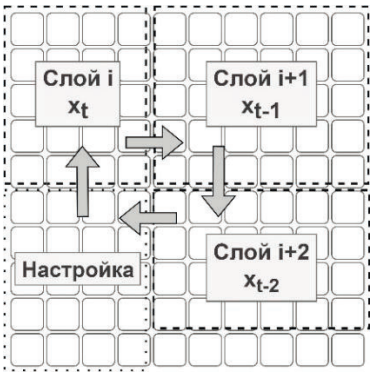

Рис. 8. Сегментированный режим функционирования среды

В данной работе предложен способ реализации рекуррентных сетей Хопфилда и LSTM на перестраиваемых ускорителях на основе вычислительных сред.

3. Реализация рекуррентной сети Хопфилда на ПВС

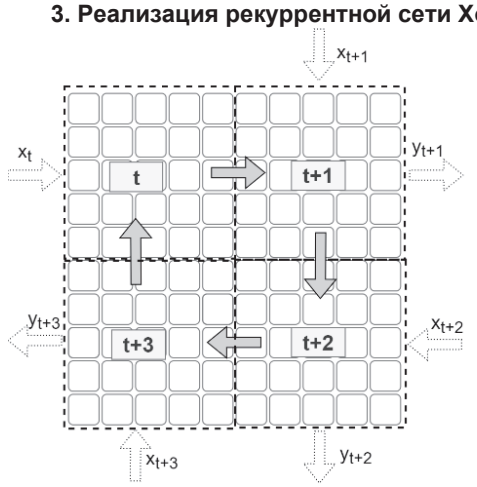

Рис. 9. Развёрнутое представление рекуррентной сети

В основе аппаратной реализации рекуррентных сетей на ПВС может использоваться их развёрнутое представление (рис. 9). Согласно такому представлению, рекуррентная сеть заменяется на глубокую нейронную сеть, каждый слой которой соответствует одному шагу работы рекуррентной сети. Все слои развернутого представления одинаковы, причём каждый эквивалентен всей рекуррентной сети. Тогда, с учётом описанного ранее сегментированного режима, рекуррентная сеть может быть реализована на среде согласно рисунку 10. Каждый сегмент среды реализует один и тот же слой, но с разным направлением передачи результата. Сигнал циклически перемещается между сегментами так долго, как требуется согласно выбранной архитектуре среды. Через внешние границы сегментов может подаваться новый входной сигнал и выводиться промежуточный результат. Так как все сегменты одинаковы, отсутствует необходимость в их перенастройке, что исключает накладные расходы и позволяет достичь максимальной скорости функционирования.

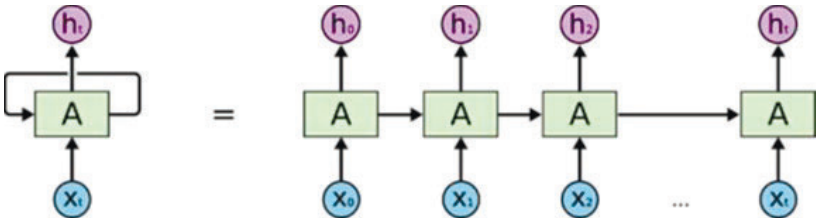

Рис. 10. Реализация развёрнутого представления рекуррентной сети на ПВС

Из рисунка 2 очевидно, что развёрнутое представление сети Хопфилда с n нейронами будет представлять собой глубокую неполносвязную сеть, каждый слой которой будет состоять из n нейронов. Сеть является неполносвязной, так как, согласно её архитектуре, каждый нейрон одного слоя не связан с соответствующим ему нейроном следующего слоя. Реализация сети Хопфилда для $n = 4$ на среде, разделённой на четыре сегмента, приведена на рисунке 11. Все сегменты среды реализуют одну и ту же конфигурацию и различаются лишь направлением передачи сигнала. Для наглядности один из сегментов выделен серым.

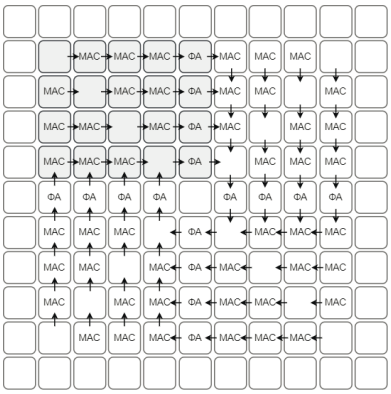

Рис. 11. Реализация сети Хопфилда с четырьмя нейронами, серым выделен один шаг развёрнутого представления сети.

Напомним, что в качестве функции активации сеть Хопфилда использует пороговую функцию (рис. 3). Несмотря на простоту, для её реализации на среде требуется решить важный вопрос – включить ли эту функцию в базис операций ВЭ, либо реализовать её на среде при помощи уже имеющихся операций. Включение функции в базис позволит затратить на её реализацию всего один ВЭ. Такое решение целесообразно, если ожидается её активное применение в реализуемых моделях. В остальных случаях более предпочтительной может быть реализация пороговой функции при помощи других операций базиса. Один из возможных вариантов реализации на пяти ВЭ представлен на рисунке 12. Он опирается на применение операции «затвор» (GAT), которая сравнивает поступающий на ключевой вход ВЭ сигнал с пороговым значением, и, при их совпадении, пропускает через себя основной сигнал (на картинке изображен слева), в ином случае на выходе элемента будет ноль. Пороговое значение затвора задано таким образом, что отпирание происходит при установленном в единицу значении знакового бита входного сигнала. Операции «источник сигнала» (SRC) формируют необходимые уровни, а «объединение» (U) и «сложение с накоплением» (MAC) осуществляют поворот и объединение сигналов.

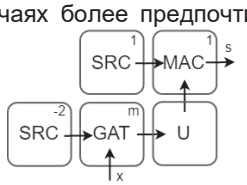

Рис. 12. Реализация пороговой функции активации на среде

4. Реализация на ПВС LSTM сети

Как было показано ранее, LSTM сеть имеет более сложное строение по сравнению с сетью Хопфилда. В связи с этим для реализации LSTM сетей нами рекомендуется разделение среды не на четыре, а на два сегмента, причём один из сегментов повёрнут относительно другого на 180° (рис. 13). Такое разбиение позволяет значительно уменьшить необходимый размер среды, что играет важную роль при реализации сетей с большим количеством нейронов и сложным состоянием ячейки (C_t).

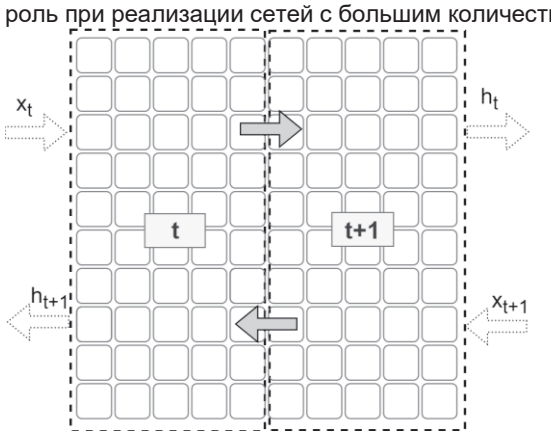

Рис. 13. Двухсегментное разбиение среды для реализации LSTM

На рисунке 4 можно видеть, что LSTM сеть включает в себя пять полносвязных слоёв (три используют в качестве активации сигмоиду, а два – гиперболический тангенс), три блока поточечного умножения, а также один блок поточечного сложения.

Как упоминалось ранее, реализация сигмоиды на ПВС (рис. 7) уже была описана в работе [V. Shatravin, 2022], где показано её высокое быстродействие. Известно, что функция активации гиперболический тангенс получается из сигмоидной при помощи несложных преобразований, благодаря чему она может быть реализована на среде аналогичным образом. Реализация блока поточечного сложения возможна при помощи группы элементов среды, настроенных на операцию умножения с накоплением с весом равным единице.

Потенциальную сложность может представлять блок поточечного умножения. В базисе операций рассматриваемой архитектуры ПВС отсутствуют операции, использование которых позволило бы получить приемлемую реализацию поточечного умножения. Решением этой проблемы может быть внедрение в базис новой операции «умножение» (MUL), которая бы выполняла соответствующее преобразование над двумя поступающими на вход элемента сигналами. Такое решение является приемлемым, так как внутри каждого элемента уже реализуется операция умножения с накоплением (MAC), которая реализуют всю необходимую логику умножения. Таким образом, внедрение операции MUL отразится лишь на маршрутизации входных сигналов внутри элемента среды, не усложняя его вычислительный блок.

В итоге благодаря небольшой модификации базиса операций элемента ПВС становится возможной реализация LSTM сетей на описанной архитектуре ПВС. Реализация одного из двух сегментов такой среды схематично представлена на рисунке 14. Цвета блоков диаграммы соответствуют участкам сети с рисунка 4. Можно видеть, что выходной сигнал предыдущего шага (h_t) выходит из среды не на предыдущем, а на текущем шаге, одновременно используясь для расчёта нового выходного значения (h_{t+1}). В то же время входной сигнал следующего шага (x_{t+1}) проходит текущий сегмент насквозь без каких-либо изменений. Такая конструкция объясняется ограничениями передачи сигнала внутри предложенной архитектуры среды. Обеспечение полной свободы перемещения сигнала внутри среды увеличило бы сложность её элементов, что приводит либо к увеличению размера требуемой среды, либо к уменьшению производительности при сохранении исходного размера. Тем не менее, такое решение не окажет значительного влияния на быстродействие модели в связи с низкими задержками передачи сигнала элементами среды (как будет показано далее, задержка передачи сигнала составляет менее 0.5 нс на один элемент среды).

5. Оценка быстродействия разработанных моделей.

Очевидно, что одной из ключевых характеристик аппаратных средств ускорения нейронных сетей является их быстродействие. В связи с тем, что на практике применяются сети разной конфигурации, целесообразно оценить быстродействие как функцию от параметров сети. В частности, от количества нейронов на разных слоях. Для этого необходимо определить зависимость от этих параметров отдельных компонентов ускорителя, оценить их быстродействие, и затем определить функциональную зависимость всей модели.

По рисункам 11 и 14 можно сделать вывод, что ключевые компоненты предлагаемых моделей это неполносвязный слой (блок операций MAC), функции активации (сигмоида, гиперболический тангенс), блоки поточечного сложения и умножения. Немаловажную роль играют также элементы, осуществляющие передачу сигнала между этими компонентами, так как передача сигнала внутри ПВС тоже занимает определённое время.

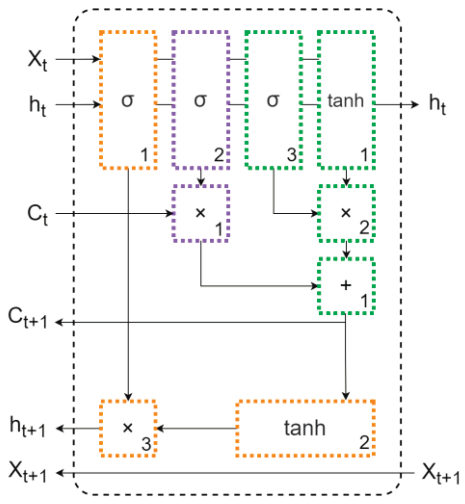

Рис. 14. Сегмент среды (правый), реализующий LSTM сеть

При выполнении оценки была проведена симуляция описанных компонентов на программируемых логических интегральных схемах (FPGA). FPGA являются широко применяемым инструментом как для исследования и прототипирования новых устройств, так и для внедрения в конечные устройства [Chang, 2016, D. Ghimire, 2022]. В рамках данного исследования была выбрана FPGA фирмы Intel Cyclone V (5CGFXC9E7F35C8), а симуляция проводилась в среде Quartus Prime (version 20.1.0, build 711 SJ Edition) при помощи встроенного пакета временных симуляций Timing Analyzer Tool. Были оценены время передачи сигнала по цепочке элементов среды разной длины, время расчёта операции умножения с накоплением (MAC) цепочкой элементов и время расчёта сигмоидной функций активации. Так как гиперболический тангенс может быть рассчитан из сигмоиды при помощи дополнительных операций умножения и сложения, его быстродействие может быть рассчитано из уже известных данных. Подробно процесс и результаты измерений описаны в [Shatravin, 2021, 2022].

Проведённые симуляции показали, что на реализацию одного элемента среды требуется 296 логических ячеек FPGA. Средняя длительность передачи сигнала через элемент менее 0.5 нс. Выполнение операции умножения с накоплением требует около 1.1 нс, а полная реализация сигмоиды 18.5 нс. Таким образом, расчёт гиперболического тангенса требует около 21.8 нс., а пороговая функция активации 2.7 нс. Данные результаты получены для элементов, работающих с 16-разрядными двоичными числами с фиксированной точкой (по 8 разрядов под целую и дробную части).

Рассмотрим модель среды, реализующую нейронную сеть Хопфилда (рис. 11). Она состоит из четырёх сегментов, каждый из которых представляет собой неполносвязный слой из n нейронов и пороговой активации. Тогда длительность наибольшего пути распространения сигнала в таком сегменте составляет:

$$t_{H1}(n) = 0.5n + 1.1(n-1) + 0.5 + 2.7. \quad 1)$$

Так как все сегменты одинаковы, то на один полный круг в среде сигналу требуется ровно в четыре раза больше времени. Если принять $n = 25$ нейронов, то $t_{H1}(25) = 42.1$ нс, а $t_{H4}(25) = 4t_{H1}(25) = 168.4$ нс. При этом потребуется среда квадратной формы с длиной стороны $2n + 1 = 51$ вычислительных элементов, то есть всего 2601 вычислительный элемент.

Рассмотрим LSTM сеть. Она имеет значительно более сложную структуру и включает в себя несколько разных полносвязных слоёв (рис. 14). Для LSTM сети можно выделить несколько ключевых параметров: размер входного сигнала (l_x), размер внутреннего состояния (l_c) и размер выходного сигнала (l_h). Четыре верхних полносвязных слоя ($\sigma_1, \sigma_2, \sigma_3, tanh_1$) принимают на вход одновременно входной сигнал и выходной сигнал с предыдущего шага, то есть длина их входного вектора составляет $l_x + l_h$. При этом их выходы имеют разный размер. Выход σ_1 имеет длину l_h, выходы $\sigma_2, \sigma_3, tanh_1$ равняются l_c. Расположенный в нижнем правом углу $tanh_2$ имеет вход длины l_c, а выход l_h. Из этих данных очевидным образом следуют размеры блоков поточечного сложения и умножения. Тогда можно видеть, что наибольшая длительность вычисления результата на каждом компоненте выражается:

$$t_{\sigma 1} = 0.5l_h + 1.1(l_h + l_x) + 18.5, \quad 2)$$

$$t_{\sigma 2} = t_{\sigma 3} = 0.5l_c + 1.1(l_h + l_x) + 18.5, \quad 3)$$

$$t_{tanh1} = 0.5l_c + 1.1(l_h + l_x) + 21.8, \quad 4)$$

$$t_{tanh2} = 0.5l_h + 1.1l_c + 21.8, \quad 5)$$

$$t_{mul1} = t_{mul2} = t_{add1} = 0.5 * 2(l_c - 1) + 1.1, \quad 6)$$

$$t_{mul3} = 0.5 * 2(l_h - 1) + 1.1\ . \quad 7)$$

В то же время надо иметь в виду, что для ПВС характерна высокая параллельность и независимое функционирование отдельных элементов в составе среды. Это значит, что формирование результата в компоненте σ_2 не ожидает завершения расчета σ_1, также как $tanh_1$ не ожидает завершения $\sigma_1, \sigma_2, \sigma_3$. Они все функционируют одновременно (с учётом задержки поступления сигналов на их входы). Из этого следует, что полная длительность расчёта результата на одном сегменте LSTM среды будет меньше суммы длительностей на всех её компонентах. Для определения фактической длительности получения результата необходимо найти и измерить критический, то есть наиболее долгий путь перемещения сигнала. Из предложенной модели среды и известных размеров её компонентов можно сделать вывод, что критический путь будет проходить слои с σ_1 по σ_3 насквозь (передача сигнала, $t_{in \to tanh}$), затем будет осуществляться расчёт компонентов $tanh_1 \to mul_2 \to add_1 \to tanh_2 \to mul_3$ с учётом задержек на конструктивно расположенных между ними элементах, осуществляющих передачу сигнала ($t_{add \to tanh2}$, $t_{tanh2 \to mul3}$). Тогда:

$$t_{LSTM1} = t_{in \to tanh1} + t_{tanh1} + t_{mul2} + t_{add1} + t_{add \to tanh2} + t_{tanh2} + t_{tanh2 \to mul3} + t_{mul3}, \quad 8)$$

$$t_{in \to tanh1} = 0.5(l_h + 2l_c), \quad 9)$$

$$t_{add \to tanh2} = 0.5(r_{tanh} + l_c), \quad 10)$$

$$t_{tanh2 \to mul3} = 0.5(|2l_c - r_{tanh}|), \quad 11)$$

где r_{tanh} – количество столбцов элементов среды, необходимых для реализации гиперболического тангенса. Примем равным 15.

Таким образом, если задать $l_x = 10$, $l_c = 25$, $l_h = 5$, то в соответствии с (8) полное время расчёта одного сегмента будет составлять:

$$t_{LSTM1} = 27.5 + 50.8 + 25.1 + 25.1 + 20 + 51.8 + 17.5 + 5.1 = 222.9 \text{ нс.} \quad 12)$$

Так как оба сегмента одинаковы, то полный круг внутри среды, соответствующий двум последовательным LSTM слоям, будет проходиться сигналом за 445.8 нс. При этом потребуется среда из 105×90 = 9450 вычислительных элементов.

Анализ альтернативных реализаций LSTM на FPGA показывает соизмеримость или преимущество предложенных моделей в отношении производительности. Исследование [He, 2021] описывает реализацию ускорителя для LSTM сети с 512 нейронами на скрытом слое (размер состояния ячейки C_t) с задержкой на обработку одного сигнала в 47.8 мкс, что при кратном масштабировании в 10.5 раз превышает полученную в данной работе оценку ускорителя на основе ПВС. [Cao, 2019] при помощи нескольких техник оптимизации (квантование данных, устранение некоторых весов и других) и параллельному расчёту поточечных операций для LSTM сети с 200 нейронами в скрытом слое удалось добиться времени выполнения 1.3 мкс, что несколько выигрывает у рассчитанных в данной работе оценок ценой уменьшения точности. Быстродействие предложенного в исследовании [Ferreira, 2016] ускорителя составляет 1.14 мкс для сети с 64 скрытыми слоями, что в 2 раза уступает предложенной нами реализации. Работа [Chang, 2016] описывает аппаратный ускоритель, требующий 932 мс на расчёт одной итерации сети со 128 скрытыми нейронами, что превышает результаты данной работы почти на 6 порядков. Следует заметить, что все рассмотренные аналоги имеют жёсткую неперестраиваемую структуру, что делает преимущества моделей на основе ПВС ещё более весомыми.

Ключевым недостатком моделей на основе ПВС является большая занимаемая площадь полупроводника. Это объясняется как накладными расходами на обеспечение перестраиваемости среды (каждый элемент среды поддерживает полный базис операций, хотя в любой момент времени выполняет лишь одну), так и высокой степенью параллельности и пространственного распределения алгоритмов. Приведённое ранее значение (296 логических ячеек FPGA на реализацию одного элемента среды) позволяет оценить количество требуемых ячеек для среды заданного размера. Уменьшение этой величины позволит увеличить производительность среды благодаря более плотному размещению её элементов на полупроводнике. Это одно из основных направлений развития исследований аппаратных ускорителей на ПВС.

Заключение

Сегодня алгоритмы машинного обучения играют важную роль во многих информационных и технических системах. Особый интерес представляют глубокие свёрточные и рекуррентные сети. Однако свойственная им высокая вычислительная сложность ограничивает их применение в широком классе маломощных мобильных и автономных систем. Одним из путей решения этой проблемы является применение аппаратных ускорителей нейронных сетей на основе перестраиваемых вычислительных сред. Такие среды обеспечивают не только высокое быстродействие благодаря распараллеливанию и пространственному распределению алгоритмов, но и высокую гибкость, что позволяет изменять выполняемые устройством алгоритмы во время его функционирования. Эта способность к динамической перенастройке отдельных участков открывает широкие возможности для разработки глубоких нейронных сетей с большим количеством слоёв.

В данной работе показано, как представление рекуррентных нейронных сетей в виде глубоких сетей может быть использовано при реализации перестраиваемого аппаратного ускорителя на ПВС. Предложены алгоритмы реализации классической рекуррентной сети Хопфилда и широко используемой сети с долгой краткосрочной памятью (LSTM). Приведены формулы оценки быстродействия LSTM сети в зависимости от её конфигурации. В частности, длительность вычисления одного шага сети с 10 входными сигналами, 25 скрытыми нейронами и 5 выходными сигналами составляет 223 нс., что значительно меньше, чем у большинства рассмотренных аналогов, к тому же лишенных способности к перенастройке. Слабым местом предложенных ускорителей на основе ПВС является большая занимаемая площадь на кристалле полупроводника. Это объясняется накладными расходами на обеспечение перестраиваемости, пространственным распределением алгоритма, высокой степенью параллельности. Так как плотность размещения элементов среды напрямую влияет на её производительность, вопрос уменьшения размера отдельного элемента будет исследоваться в рамках дальнейших работ. Несмотря на эту особенность, аппаратные ускорители на основе ПВС демонстрируют высокие показатели по ключевым параметрам, что подтверждает перспективность их дальнейшего исследования.

Благодарности:

Исследование выполнено за счет гранта Российского научного фонда
№ 21-71-00012, https://rscf.ru/project/21-71-00012/

Литература

1. Шатравин В., Шашев Д. В. Разработка алгоритма настройки перестраиваемой вычислительной среды в составе аппаратного ускорителя искусственных нейронных сетей / В. Шатравин, Д. В. Шашев // Цифровая экономика. – 2022. – 20(4). – с. 11–18.

2. Cao S. Efficient and Effective Sparse LSTM on FPGA with Bank-Balanced Sparsity / S. Cao [et al.] // Proceedings of the 2019 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays. – 2019. – pp. 63-72.
3. Chang A. X. M. Recurrent Neural Networks Hardware Implementation on FPGA / A. X. M. Chang, B. Martini, E. Culurciello // arXiv:1511.05552v4. – 2016. – pp. 1-7.
4. Chen Y. H. Eyeriss v2: A flexible accelerator for emerging deep neural networks on mobile devices / Y. H. Chen [et al.] // IEEE Emerging and Selected Topics in Circuits and Systems (Jetcas). – 2019. – 9. – pp. 292-308.
5. Ferreira J. C., Fonseca J. An FPGA Implementation of a Long Short-Term Memory Neural Network / J. C. Ferreira, J. Fonseca // International Conference on ReConFigurable Computing and FPGAs (ReConFig). – 2016. – pp. 1-8.
6. Ghimire D., Kil, D., Kim, S.-h. A Survey on Efficient Convolutional Neural Networks and Hardware Acceleration // J. Electronics, MDPI. – 2022, 11. – vol. 945. – pp. 1-23.
7. He D. An FPGA-Based LSTM Acceleration Engine for Deep Learning Frameworks / D. He [et al.] // Electronics. – 2021. – 10(6). – pp. 1-15.
8. Hochreiter S., Schmidhuber J. Long Short-term Memory / S. Hochreiter, J. Schmidhuber // Neural computation. – 1997. – 9(8).
9. Shatravin V., Shashev D., Shidlovskiy S. Sigmoid Activation Implementation for Neural Networks Hardware Accelerators Based on Reconfigurable Computing Environments for Low-Power Intelligent Systems // MDPI: Applied Sciences. – 2022. – 12(10).
10. Shatravin V., Shashev D. V., Shidlovskiy S.V. Applying the Reconfigurable Computing Environment Concept to the Deep Neural Network Accelerators Development // International Conference on Information Technology (ICIT) – 2021. – pp. 842-845.
11. Understanding LSTM Networks. Colah's blog [Электронный ресурс]. – URL: https://colah.github.io/posts/2015-08-Understanding-LSTMs/ (дата обращения 12.10.2022).

References in Cyrillics

1. Shatravin V., Shashev D. V. Development of configuration algorithm for reconfigurable computing environment in neural networks hardware accelerators / V. Shatravin, D. V. Shashev // Digital Economy. – 2022. – 20(4). – pp. 11-18.

Шатравин Владислав, аспирант, млад. науч. сотр. ТГУ, Томск
(shatravin@stud.tsu.ru)
Шашев Дмитрий Вадимович, к.т.н, доцент ТГУ, Томск
(dshashev@mail.tsu.ru)

Ключевые слова
Рекуррентные нейронные сети, LSTM, перестраиваемые вычислительные среды, реконфигурируемые аппаратные ускорители.

Vladislav Shatravin, Acceleration of recurrent neural networks with computing environments

Keywords
Recurrent neural networks, LSTM, reconfigurable computing environments, reconfigurable hardware accelerators.
DOI: 10.34706/DE-2023-01-04
JEL: C63 – Вычислительные методы, моделирование.

Abstract
The application of modern machine learning algorithms in technical systems is limited by the hardware they use. The problem is particularly serious when using large neural networks in low-power and autonomous systems that have severe weight and power consumption restrictions. Majority of modern neural network hardware accelerators either have high both power consumption and weight or they are highly specialized for a small set of algorithms. One possible solution to the problem is the use of dynamically reconfigurable hardware accelerators that can change the implemented algorithms at run time. The accelerators can be based on the principles of reconfigurable computing environments (RCE). This paper presents implementations of Hopfield and long short-term memory (LSTM) recurrent networks on RCE-based accelerators. The performance evaluations of the developed models were determined through simulations on FPGA. Estimates show the high performance of the presented models in comparison with analogues, however, the requirements for the area on a chip are also higher. According to estimates, an LSTM network with 25 hidden neurons will be calculated in 223 ns. The results obtained allow to conclude that there is a high potential for using RCE-based accelerators for recurrent networks and the need for further optimization.