УДК 004.75

ВЫЧИСЛЕНИЕ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ НА ВЫЧИСЛИТЕЛЬНЫХ КЛАСТЕРАХ ИЛИ ЛВС

М. С. Герасименко

ОАО «Концерн «Созвездие», г. Воронеж

Поступила в редакцию 01.03.2010 г.

Аннотация. Рассмотрена теоретическая возможность и целесообразность использования распределенных вычислений для сокращения времени расчета искусственных нейронных сетей (ИНС) на вычислительных кластерах и в локальных вычислительных сетях (ЛВС). Проведено сравнение с методом команды экспертов и методом виртуальных частиц.
Ключевые слова: распределенные вычисления, кластер, искусственные нейронные сети (ИНС).

Abstract. Theoretical possibility and practicability of distributed computing artificial neural nets on compute clusters or in a local area networks (LAN) for decrease time of it calculation are offered. Comparison with expert group method and virtual particles method are realized.
Key words: distributed computing, cluster, artificial neural net (ANN).

ВВЕДЕНИЕ

Искусственные нейронные сети стремительно развиваются, размер нейронных сетей быстро растет, а следовательно, возрастает и объем вычислений. С увеличением размеров сети требуется увеличение вычислительных мощностей. С точки зрения развития вычислительной техники и программирования, нейронная сеть — способ решения проблемы эффективного параллелизма, однако этот параллелизм является «мелкозернистым», что затрудняет реализацию вычислений. Существует несколько основных способов реализации искусственных нейронных сетей. Наиболее эффективным из них является создание нейрокомпьютеров, аппаратная составляющая которых полностью основана на нейропроцессорах. Другим типом устройств являются нейроускорители, реализованные в виде карты или модуля с распараллеливанием операций на аппаратном уровне. Несмотря на свои впечатляющие возможности, такие устройства не очень распространены на компьютерном рынке из-за высокой цены (от единиц до десятков тысяч долларов для карт и от десятков до сотен тысяч долларов для модулей), а также специфики и сложности освоения. Наибольшее распространение получила эмуляция процессов, происходящих в нейронных сетях, на ЭВМ, однако в этом случае вычисление ИНС происходит последовательно, и не приходится говорить о каком-либо распараллеливании процесса вычислений [1, 2].

Искусственные нейронные сети представляют собой связевые системы. В отличие от цифровых микропроцессорных систем, представляющих собой сложные комбинации процессорных и запоминающих блоков, нейропроцессоры содержат память, распределенную в связях между очень простыми процессорами. Тем самым основная нагрузка ложится на архитектуру системы, детали которой, в свою очередь, определяются межнейронными связями [2, 3].

С точки зрения реализации 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. Каждый процессор подобной сети имеет дело только с сигналами, которые он периодически получает, и сигналами, которые он посылает другим процессорам [3].

Типичный пример сети прямого распространения показан на рис. 1. Нейроны регулярным образом организованы в слои. Входной слой — распределительный — служит просто для ввода значений входных переменных. Каждый из скрытых и выходных нейронов соединен со всеми элементами предыдущего слоя. Можно

было бы рассматривать сети, в которых нейроны связаны только с некоторыми из нейронов предыдущего слоя; однако, для большинства приложений сети с полной системой связей между слоями предпочтительнее [3].

Идея распараллеливания вычислений базируется на том, что большинство задач может быть разделено на набор меньших задач, которые могут быть решены одновременно на нескольких вычислительных узлах. Частным случаем параллельных вычислений являются распределённые вычисления, которые представляют собой способ решения трудоёмких вычислительных задач с использованием двух и более компьютеров, объединённых в сеть. Ориентация на использование в вычислительных системах серийных микропроцессоров и памяти позволяет снизить стоимость, приходящуюся на единицу производительности [4, 5].

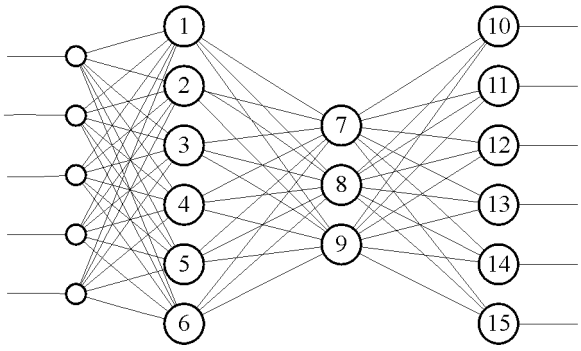


Рис. 1. Пример сети прямого распространения

Общей особенностью масштабируемых систем является распределенность процессов обработки и распределенность данных. С одной стороны в распределении вычислений и данных заложена возможность повышения производительности таких систем. С другой стороны интенсивный обмен данными, часто и не вызванный информационными зависимостями между процессами обработки, сильно загружает коммуникационную среду. Естественно, это ограничивает возможности масштабирования систем. Следовательно, эффективность выполнения программ в масштабируемых архитектурах в значительной мере зависит от организации взаимодействия распределенных процессов. Для организации распределенных вычислений необходимо, чтобы решаемая задача была сегментирована — разделена на подзадачи, которые могут вычисляться параллельно. Однако, не всякую задачу можно «распараллелить» и ускорить её решение с помощью распределенных вычислений [5].

Обработка данных искусственной нейронной сетью и особенно ее обучение, осуществляемое за большое число итераций, требует значительных вычислительных мощностей. В настоящее время широкому кругу пользователей стали доступны мощные параллельные вычислительные системы (ПВС), собираемые из элементов высокой степени готовности, — кластеры. Алгоритмы обучения и функционирования ИНС изначально являются параллельными, но этот параллелизм «мелкозернистый», т.е. объем вычислений в отдельном узле сети невелик, а количество связей между узлами и интенсивность обмена значительны. По этой причине данные алгоритмы в исходном виде плохо подходят для реализации на большинстве ПВС [6]. Для эффективной реализации на кластере требуется адаптация модели ИНС к особенностям его архитектуры. Для сокращения времени обучения ИНС предложены метод виртуальных частиц и команда экспертов [6].

Оба эти метода позволяют сократить время, необходимое на обучение нейронной сети, однако, при работе уже обученной ИНС время обработки результатов, несмотря на многократное увеличение вычислительных мощностей, напротив, увеличится. Для команды экспертов это вызвано тем что для каждого входа необходимо вычислять выходные значения сетей для всей команды экспертов, поскольку окончательное решение принимается на основе всей совокупности ИНС, составляющих команду экспертов, а это, естественно, невозможно осуществить быстрее чем произвести вычисление одной из ИНС на самом медленном из узлов, при предложенном способе реализации. Метод виртуальных частиц позволяет сократить время получения одной обученной нейронной сети, которую впоследствии можно использовать на одном из узлов предложенной кластерной системы, однако все равно не позволяет сократить время вычисления результата работы этой ИНС методом распределения вычислений.

Использование мелкозернистой параллельности нейронных сетей для проведения распределенных вычислений считается неоправданным, поскольку число связей и узлов сети очень велико, а объем вычислений в каждом из узлов сети незначителен. В частности, для сети, изоб-

раженной на рис. 1 и насчитывающей всего 15 узлов, число связей оказывается равным 66 и стремительно растет с увеличением размеров сети. Для вычисления сети, содержащей всего 1000 нейронов, потребуется вычислительная сеть, содержащая 1000 узлов, и, в зависимости от структуры сети, около 250000 связей. При этом эффективность работы такой сети будет значительно ниже, чем при вычислении ИНС на одном из узлов распределенной вычислительной системы, поскольку потребуется значительное время для передачи данных между сетями, ожидание этих данных и синхронизацию вычислений.

В настоящей работе рассматривается возможность и целесообразность использования кластерных параллельных вычислительных систем для ускорения вычисления ИНС с применением метода распределения фрагментов нейронной сети между узлами вычислительной системы.

АНАЛИЗ ВОЗМОЖНОСТЕЙ И ЦЕЛЕСООБРАЗНОСТИ ИСПОЛЬЗОВАНИЯ КЛАСТЕРНЫХ ПАРАЛЛЕЛЬНЫХ ВЫЧИСЛИТЕЛЬНЫХ СИСТЕМ ДЛЯ УСКОРЕНИЯ ВЫЧИСЛЕНИЯ ИНС

Особенности архитектуры ИНС позволяют выделить фрагменты нейронной сети, которые могут быть вычислены локально, без обмена информацией с другими узлами вычислительной системы. Именно это позволяет распределить фрагменты нейронной сети таким образом, чтобы минимизировать сетевой обмен и увеличить объем вычислений в каждом из узлов вычислительной системы. Распределение между вычислительными центрами можно провести множеством способов. Один из вариантов такого разбиения представлен на рис. 2а. Нейронная сеть в данном случае распределена по пяти вычислительным узлам, каждый из которых обрабатывает три нейрона ИНС. Если упростить схему до уровня узлов вычислительной сети, получим схему, представленную на рис. 2б, структура которой не отличается от структуры ИНС, только входными и выходными значениями являются не скалярные величины, а векторы.

Такое разбиение, естественно, приводит к уменьшению количества узлов распределенной вычислительной сети до необходимого количества. Кроме того, возрастает объем вычислений, приходящихся на один узел, уменьшается количество связей и сокращается объем передаваемой по сети информации. Именно это и позволяет использовать распределенные вычисления, поскольку для эффективного использования ПВС требуется, чтобы время вычисления узлов вычислительной сети было больше времени, необходимого для передачи информации между вычислительными центрами.

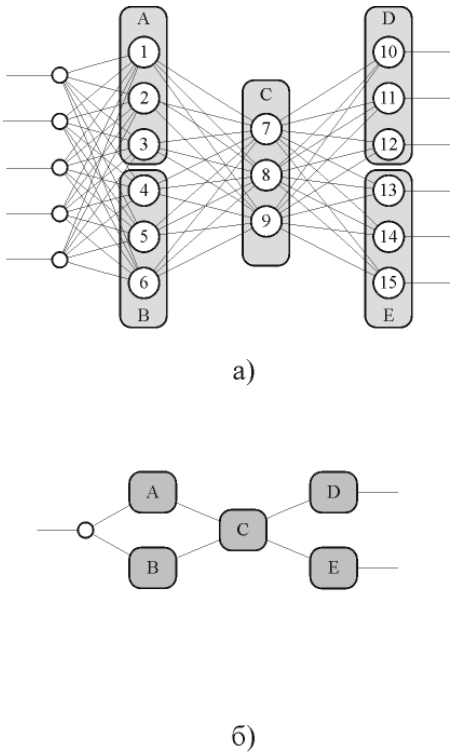
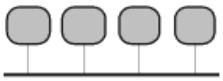


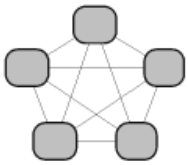
Рис. 2. а) один из способов распределения нейронной сети между пятью вычислительными узлами; б) схема, упрощенная до уровня узлов вычислительной сети и связей между ними

При фиксированном числе узлов ПВС увеличение размеров вычисляемой нейронной сети приводит к значительному увеличению объема вычислений, приходящихся на каждый узел, и небольшому увеличению объемов передаваемой по сети информации. Число связей между узлами вычислительной сети при этом остается неизменным. Все это приводит к увеличению эффективности использования распределенных вычислений при увеличении размеров ИНС за счет сокращения времени ожидания информации и более полного использования вычислительных мощностей ПВС.

Эффективность и метод реализации алгоритма существенным образом зависит от способа организации передачи информации между узлами. Рассмотрим четыре основных варианта, схематично изображенных на рис. 3:

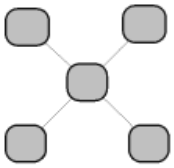


а)



б)

в)



г)

Рис. 3. Возможные схемы организации передачи данных в сети. а) используется общая среда передачи данных (т.е. любое передаваемое сообщение получат все вычислительные узлы, и в один момент времени передавать информацию может только один из узлов сети); б) полный граф, когда все узлы сети соединены независимыми линиями связи; в) вычислительные узлы объединены в сеть с помощью коммутаторов или маршрутизаторов; г) соединение типа «звезда»

Рассмотрим все эти архитектуры с точки зрения реализации рассматриваемого алгоритма. Как видно из схемы, приведенной на рис. 2б, результат вычисления каждого из узлов распределенной вычислительной сети передается нескольким следующим узлам. Таким образом, использование общей среды передачи данных (рис. 3а) позволяет не дублировать передачу результатов вычисления всем узлам, для которых оно предназначено. Однако этот способ организации передачи данных имеет и свои недостатки: все узлы вычислительной сети должны тратить вычислительные ресурсы на фильтрацию пакетов, предназначенных другим узлам. Поскольку одновременно передавать информацию может только один из узлов, все остальные узлы, имеющие подготовленные для передачи данные, должны ожидать момент, когда будет возможен выход на передачу. Такой способ передачи данных накладывает ограничения на возможность увеличения числа узлов распределенной вычислительной сети.

Архитектура полного графа (рис. 3б) позволяет передавать данные только узлам, для которых они предназначены. Кроме того, не возникает проблем, связанных с одновременным выходом на передачу разных узлов сети. Из недостатков такого подхода можно отметить, что каждому узлу необходимо передавать одну и ту же информацию многократно — для каждого из узлов, которым она предназначена, если это не реализовано аппаратно. При таком способе организации сети очень сложно создавать распределенные вычислительные системы с большим количеством узлов, поскольку с увеличением количества узлов вычислительной сети количество используемых связей быстро возрастает.

Сеть, организованная с помощью коммутаторов или маршрутизаторов (рис. 3в), имеет много общего с предыдущим, но рассылка сообщений множеству узлов вычислительной сети осуществляется маршрутизаторами. Благодаря этому многократно сокращается количество физических линий связи, а основная нагрузка по доставке сообщений ложится на специализированные устройства.

Архитектура типа «звезда» (рис. 3г), хотя и является очень распространенным при организации распределенных вычислений, плохо подходит для вычисления нейронных сетей, поскольку нет явно выделенного сервера, фор-

мирующего фрагменты задач для вычисления на других узлах вычислительной системы и обобщающего результаты. В итоге подобная структура организации передачи данных приводит к значительному увеличению трафика, а следовательно, уменьшению эффективности функционирования вычислительной сети.

Естественно, представленный на рис. 2 метод разбиения не является единственным. Рассмотрим вариант, представленный на рис. 4. В этом случае для вычисления 7-го нейрона узла A необходимы результаты вычисления нейронов 3 и 4, принадлежащих узлу B. Это приводит к тому, то узлы A, B и C не могут быть вычислены локально, что необоснованно усложняет задачу, увеличивает объем передаваемой по сети информации и, в конечном итоге, увеличивает время вычисления ИНС.

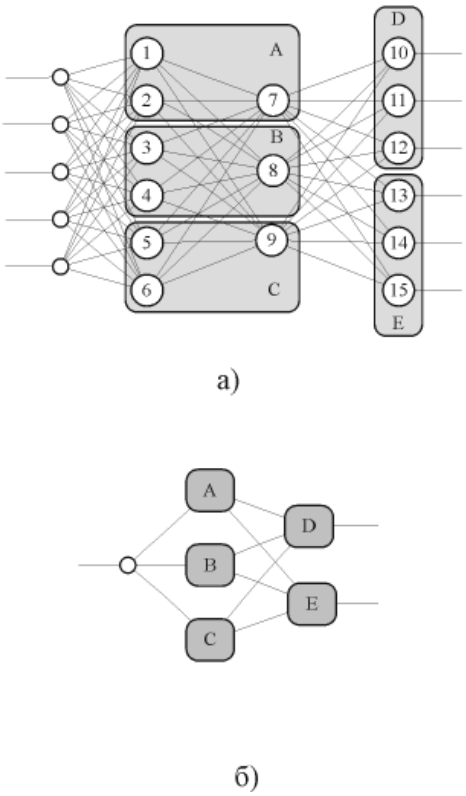


Рис. 4. а) один из способов распределения нейронной сети между пятью вычислительными узлами; б) схема, упрощенная до уровня узлов вычислительной сети и связей между ними.

Поскольку для вычисления нейронов, расположенных в промежуточных и выходном слое нейронной сети, требуется результат вычисления всех предыдущих слоев, существует потребность в контролировании получения всех данных, необходимых для вычисления узла, и принадлежности этих данных к одному и тому же циклу работы ИНС. Кроме того, поскольку разные узлы осуществляют вычисление своего участка нейронной сети за различное время, для оптимизации процесса вычислений целесообразно использовать входные и выходные буферы данных, а в протоколе обмена должна быть предусмотрена возможность передачи информации о переполнении и освобождении буферов приема. При переполнении буферов передачи вычисления в сети придется приостанавливать и ожидать обработки данных последующими узлами. В любом случае выходное значение всей нейронной сети не может быть вычислено быстрее, чем производится вычисление самого медленного из узлов распределенной вычислительной системы.

Наибольшая эффективность работы достигается при равенстве времени вычисления каждым из узлов вычислительной системы своего участка нейронной сети. Определение быстродействия каждого из узлов вычислительной системы возможно на основании времени, затраченного на вычисление тестовой нейронной сети. Эти признаки и знание структуры всей нейронной сети возможно положить в основу алгоритма распределения ИНС между узлами вычислительной системы. Возможность автоматического распределения нейронной сети между узлами вычислительной системы является несомненным достоинством систем подобного типа, поскольку, в отличие от большинства других способов, организации вычислений, не требует специальных знаний от пользователя.

Рассмотренный вариант многослойной архитектуры ИНС является наиболее распространенным, но встречаются и другие архитектуры. Самым сложным является вариант полносвязной ИНС. В этом случае обмен информацией между узлами вычислительной сети будет наиболее интенсивным. Тем не менее, даже в таком случае рассмотренный подход оказывается применимым. Для случаев слабосвязанных ИНС (когда существуют большие участки ИНС с малым количеством связей между ними) рассмотренный метод является наиболее эффективным, поскольку возможно существенно снизить интенсивность обмена между узлами распределенной вычислительной сети.

ВЫВОДЫ

1. Распределение искусственной нейронной сети между несколькими вычислительными центрами применимо только для достаточно больших ИНС, когда время вычисления каждого из узлов распределенной вычислительной системы значительно больше времени передачи информации между ними. Для малых ИНС эффективнее прямое вычисление сети на одном из узлов вычислительной системы.

2. Существует возможность автоматизировать процесс распределения нейронной сети между вычислительными узлами ПВС, что упрощает работу с подобными вычислительными сетями. Широкая доступность ЭВМ, объединенных локальной сетью, или специализированных кластерных ПВС, а также простота использования позволяет сделать подобные системы вычисления нейронных сетей доступными широкому кругу пользователей.

3. В отличие от методов виртуальных частиц и команды экспертов для вычисления искусственной нейронной сети не используются никакие надстройки, вводящие дополнительные ИНС и приводящие к необходимости учитывать и анализировать все полученные результаты. Следовательно, возможно использовать все алгоритмы, разработанные для искусственных нейронных сетей в исходном виде, не приспосабливая их к используемой модели вычислений.

СПИСОК ЛИТЕРАТУРЫ

1. *Круг П. Г.* Нейронные сети и нейрокомпьютеры. / П. Г. Круг — МЭИ, 2002.

2. *Колесников.* Аппаратная реализация нейронных сетей / Колесников — «КОМПЬЮТЕР-ИНФОРМ» № 17, 2005.

3. *Уоссермен Ф.* Нейрокомпьютерная техника. Теория и практика. / Ф. Уоссермен — пер. с англ. 1992.

4. *Корнеев В. В.* Параллельные вычислительные системы. / Корнеев В. В — Москва, 1999.

5. *Топорков В. В.* Модели распределенных вычислений. / В. В. Топорков — М.: ФИЗМАТЛИТ, 2004.

6. *Фомин Ю. Н., Сиротинина Н. Ю.* Исследование эффективности реализации методов виртуальных частиц и команды экспертов на кластерной архитектуре. / Ю. Н. Фомин, Н. Ю. Сиротинина — «Наука и образование. Инженерное образование» № 2 февраль 2007.

Герасименко Максим Сергеевич – инж.-констр., ОАО «Концерн «Созвездие», г. Воронеж. Тел. 8-(905)-049-43-50. E-mail: grsms@mail.ru

Gerasimenko M. S. – design engineer, JSC «Sozvezdie» Concern». E-mail: grsms@mail.ru