 DOI 10.25699/SSSB.2019.2(26).32523

05.13.18

ПАРАЛЛЕЛЬНАЯ РЕАЛИЗАЦИЯ МНОГОСЛОЙНОЙ НЕЙРОННОЙ СЕТИ С ЭЛЕМЕНТАМИ САМООБУЧЕНИЯ НА ГЕТЕРОГЕННОМ КОМПЬЮТЕРЕ

А.А. Малявко

Новосибирский государственный технический университет, г. Новосибирск

В статье рассматриваются возможные способы уменьшения затрат времени на симуляцию искусственной нейронной сети, архитектура которой ориентирована на изучение механизмов самообучения. На основе сопоставления с биологическими нейронными системами, заведомо способными к самообучению, формулируются некоторые предположения о возможной структуре такой сети в виде совокупности нескольких функционально разнотипных многослойных блоков нейронов. Связи между нейронами направлены преимущественно от входа сети к ее к выходу, но имеются и связи между нейронами одного слоя, а также связи обратной направленности. Эффект самообучения возможно, будет достигнут при реализации непрерывного циклического моделирования работы сети, что полностью соответствует механизмам функционирования биологических прототипов. Непрерывная симуляция сети с большим количеством нейронов требует очень больших затрат компьютерного времени. Поэтому актуальной является ориентация на использование гетерогенных компьютеров, предоставляющих значительно большие вычислительные мощности по сравнению с компьютерами традиционной архитектуры. Описывается параллельная программная модель, разработанная для проведения экспериментов по изучению механизмов самообучения на многоядерных компьютерах с несколькими графическими процессорами, и реализованный в этой модели алгоритм распределения и балансировки нагрузки графических процессоров и ядер центрального процессора. Приводятся результаты экспериментов на двух различных гетерогенных компьютерах, показывающие сравнительно слабый эффект ускорения вычислений за счет использования нескольких GPU. Этот эффект можно объяснить необходимостью постоянного перемещения больших объемов данных между основной памятью и памятью графических процессоров вследствие непрерывного переконфигурирования параметров межнейронных связей, осуществляемого симулятором при исследовании алгоритмов самообучения.

Ключевые слова: нейронная сеть, самообучение, гетерогенный компьютер, графический процессор.

ВВЕДЕНИЕ

Известно, что биологические нейронные системы любых живых организмов в той или иной степени способны к самообучению [1, 2]. Под самообучением будем понимать выработку поведения – некоторой типовой последовательности действий в ответ на однотипные последовательности воздействий окружающей среды [3, 4]. Большую роль в самообучении молодых организмов имеют разнообразные игры. Поэтому исследования с целью создания самообучающейся искусственной нейронной сети предполагается выполнять путем имитации ее игры с учителем. Конкретный вид игры особого значения не имеет. Важно, что игра – это чередование «ходов» участников. Последовательность ходов каждого участника имеет конкретную цель – выигрыш в сеансе игры. Поэтому вся последовательность ходов в момент окончания сеанса может быть оценена как плохая или как хорошая с точки зрения достижения цели. Соответственно этой общей оценке может быть оценен и каждый ход последовательности, но только апостериори.

Для реализации самообучения на этой основе биологические нейронные системы, вероятно, обладают рядом свойств и механизмов, таких как:

–память, способная хранить (в точном или приближенном виде) последовательность ходов игры и условия, в которых эти ходы были сделаны;

–способность «прокручивать» в прямом и/или обратном порядке ходы из памяти;

–способность оценивать свои ходы и ходы противника на основе итоговой оценки результата данной игры;

–память, хранящая оценки ранее сделанных ходов – как собственных, так и ходов противника;

–способность сопоставлять текущие условия с хранимыми условиями и оценками сделанных в этих условиях ходов для выработки нового хода при реализации последующего игрового сеанса.

На данный момент нет точных данных о том, каким образом биологическая нейронная система (далее – мозг) в процессе своего развития формирует структуры, вырабатывающие поведение, адекватное воздействию окружающей среды, т.е. реализующие вышеперечисленный функционал. Достаточно ясно только то, что отдельные нейроны и целые их ансамбли играют собственные специфические роли в процессе непрерывного функционирования и самообучения [3]. Некоторые из них осуществляют распознавание входных сигналов, другие – сопоставление с хранимыми образцами, третьи – принятие решений и т.д. О том, как формируются такие нейронные ансамбли, могут быть высказаны только более или менее обоснованные предположения [5, 6]. Поэтому представляется целесообразным исследование воз-

можных вариантов реализации механизмов самообучения искусственных нейронных сетей [4]. Эти исследования связаны с формированием объемных специфических нейросетевых структур, моделирование поведения которых требует очень больших затрат компьютерного времени. Поэтому оказывается целесообразным использование параллельных вычислителей [6 – 8]. В качестве таких вычислителей могут использоваться гетерогенные многоядерные компьютеры с несколькими графическими процессорами.

ОСНОВНАЯ ЧАСТЬ

Программная модель нейронной сети обучения разработана для выявления и исследования возможных механизмов самообучения искусственной нейронной сети игровому поведению [5]. Взаимодействие основных составляющих частей программной модели нейронной сети друг с другом и с окружающей средой организовано по классической схеме и показано на рис. 1. Предполагается, что роль окружающей среды реализует противник по сеансу игры, он же – учитель.

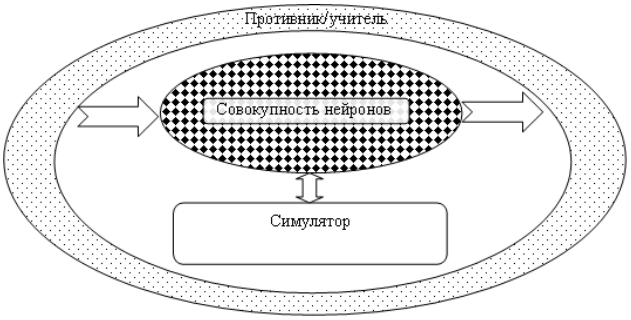

Рис. 1. Схема взаимодействия элементов программной модели

Противник/учитель взаимодействует только с совокупностью нейронов сети, т.е. формирует входные данные и получает выходные. Симулятор, реализующий алгоритмы модификации параметров нейронной сети (самообучения), извлекает всю необходимую для этого информацию из состояний некоторых ее нейронов и напрямую с учителем не взаимодействует. Предполагается, что симулятор выявляет такие нейроны, которые далее будут называть центрами эмоциональной самооценки [9,10], в процессе реализации игры на основании сопоставления «ходов» игроков с достигнутыми в игре результатами. Предполагается также, что по аналогии с биологическим прототипом работа совокупности нейронов моделируется циклически шаг за шагом независимо от активности окружающей среды.

Совокупность нейронов сети представляет собой функционально неоднородный набор связанных многослойных блоков, слои которых также имеют различную функциональную нацеленность. Один из возможных примеров блочно-слоевой структуры сети приведен на рис. 2. Эту структуру можно рассматривать, например, как аналог части мозга, в которой звуковые сигналы от одного из парных органов чувств (глаз или ухо) подвергаются первичной обработке рабочими блоками 1 и 2. После этого рабочий блок 3 преобразует два потока сигналов в один объемный и отправляет на дальнейшую обработку в высшие отделы коры. Каждый блок представляет собой совокупность нескольких слоев, на рис. 2. слои внутри блоков разделены пунктирными линиями. Внутри блоков межнейронные связи преимущественно ориентированы слева направо (от входных слоев к выходным), но могут существовать и встречно направленные связи, а также связи между нейронами внутри слоя.

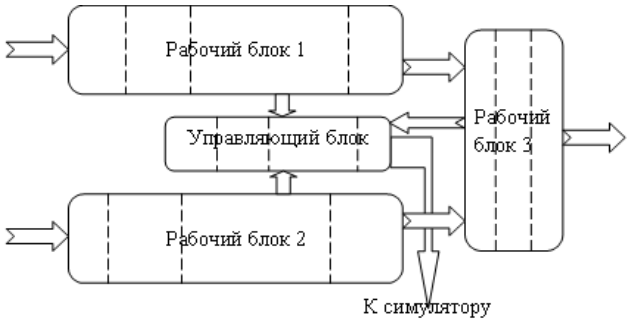

Рис. 2. Блочно-слоевая структура совокупности нейронов

Сформулированные предположения о механизмах работы мозга при самообучении и стремление добиться этого эффекта в модели привели к принятию следующих решений при начальном конфигурировании совокупности нейронов сети. В момент запуска программной модели симулятор считывает параметры структуры сети из конфигурационного файла и формирует внутренние структуры данных. В связи с тем, что точная структура самообучаемой сети на данный момент неизвестна и в силу того, что количество нейронов в ней предположительно должно составлять величину порядка нескольких сотен тысяч и более, исследователь не определяет каждый синапс в конфигурационном файле. В нем указываются количества нейронов в каждом слое и диапазоны значений параметров межнейронных связей (сколько в среднем синапсов имеет нейрон из некоторого слоя и каковы в среднем веса этих синапсов). На основании этих данных синапсы нейронов формируются симулятором случайным образом. Сформированная таким образом структура сети может быть сохранена на диске для последующего восстановления и продолжения работы с нею.

Блоки нейронов в модели могут быть рабочими и управляющими. Рабочих блоков может быть несколько, управляющих – один. Рабочие блоки воспринимают воздействия окружающей среды или выходы других блоков и формируют выходные воздействия для окружающей среды или других блоков. Управ-

ляющий блок предназначен для отслеживания состояний нейронов рабочих блоков и сопоставления выработанной сетью реакции на входное воздействие с тем, как ее оценивает учитель. Для каждого блока задаются диапазоны значений порогов срабатывания нейронов и весов межнейронных связей. Эти диапазоны, установленные для блока, наследуются составляющими его слоями в том случае, если для слоя не заданы его собственные одноименные диапазоны.

Слои блока могут быть определены как входные (input), внутренние обрабатывающие (inner), эмоциональные (emotion) или выходные (output). Функциональным назначением нейронов эмоционального слоя является выработка самооценки сети для симулятора. Для каждого слоя в конфигурационном файле задается количество нейронов. Для каждого блока и для каждого слоя могут быть заданы диапазоны значений порогов срабатывания нейронов и весов межнейронных связей. Если какой-либо из этих диапазонов для слоя не задан, то он наследуется из содержащего этот слой блока. Синапсы нейронов каждого слоя описываются путем задания совокупностей, состоящих из номера блока, номера слоя, диапазона ширины пучка связей, идущих из этого слоя, и опционально – диапазона весов.

В результате обработки конфигурационного файла формируются все нейроны и межнейронные связи. Случайным образом формируются начальные состояния всех нейронов. Для каждого нейрона начальное состояние получает случайное значение –1 или 0 или 1 (вероятность каждого значения 1/3).

Далее симулятор программной модели циклически реализует пошаговую имитацию работы сети. На каждом шаге выполняется два этапа:

– пересчет состояний всех нейронов сети;

– обработка полученной совокупности и, возможно, формирование ответа сети на входное воздействие, а также, возможно, модификация порогов срабатывания некоторых нейронов и/или весов некоторых межнейронных связей.

Симулятор может работать последовательно на одном ядре центрального процессора или использовать все ядра центрального процессора и/или один или несколько графических процессоров.

Вся совокупность данных, хранящих состояния нейронов сети и их связей, в программной модели разбита на несколько массивов, первоначально размещаемых в основной памяти компьютера. При параллельной обработке часть этих данных копируется в память графического процессора (или нескольких процессоров) для реализации на нем (на них) этапа пересчета состояний. Этап обработки симулятором реализуется всегда на ядрах центрального процессора. Структура совокупности массивов данных была спроектирована так, чтобы обеспечить реализацию функций исследуемой сети, минимизировать объемы пересылок данных между разными уровнями памяти и реализовать динамическое перераспределение вычислительной нагрузки на основе измеряемых на каждом шаге затрат времени каждого элемента гетерогенного компьютера. Совокупность обрабатываемых данных содержит следующие массивы.

1. Массив структур *Neurons*, каждая из которых содержит:
   1.1 Порог срабатывания нейрона
   1.2 Номер первого синапса
   1.3 Номер последнего синапса
2. Массив синапсов *Synapses*. Каждый элемент этого массива содержит номер нейрона, выход которого связан со входом данного нейрона. Все синапсы одного нейрона расположены в этом массиве последовательно.
3. Массив весов межнейронных связей *Weights*. Каждая межнейронная связь или синапс однозначно определяется номером нейрона и весом.
4. Два массива значений выходов нейронов *States0* и *States1* для четных и нечетных шагов. Перед шагом работы сети с четным номером актуальными по входам являются значения из массива *States0*, формируемые на этом шаге новые состояния нейронов сохраняются в массиве *States1*. Соответственно, перед шагом с нечетным номером входные значения находятся в массиве *States1*, а вырабатываемые сохраняются в массиве *States0*.
5. Массив флажков *Flags*, формируемых симулятором по результатам обработки множества состояний после каждого шага и определяющих способ последующего использования нейрона.
6. Три массива счетчиков срабатываний нейронов *Counts<i>* (торможений, срабатываний и отсутствия срабатываний), используемых симулятором для формирования значений флажков.

Если моделирование сети выполняется только на ядрах центрального процессора, то все эти массивы располагаются в его основной памяти.

Если используется один графический процессор, то в его память копируются полностью массивы *Neurons*, *Synapses*, *Weights*, *States0* и *States1*. После каждого шага пересчета состояний нейронов один (актуальный) из массивов *States0* или *States1* копируется в основную память. Массивы *Flags* и *Counts<i>* формируются и обрабатываются симулятором в основной памяти.

При использовании нескольких графических процессоров в память каждого из них копируются целиком массивы *Synapses* и *Weights*. Массивы *Neurons*, *States0* и *States1* распределяются между используемым устройствам в начальный момент одинаковыми частями. Далее на каждом шаге измеряются затраты времени каждого графического процессора, вычисляется среднее значение и пропорционально отклонению от него для каждого устройства массивы перераспределяются.

Программная модель написана на языке C++ с использованием OpenMP и OpenCL и с помощью механизма событий обеспечивает одновременный запуск всех выбранных исследователем устройств гетерогенного компьютера на обработку выделенных им данных. При наличии хотя бы одного графического процессора выполнение этапа обработки состояний нейронов, полученных на некотором шаге, симулятором во времени совмещается с выполнением этапа пересчета их состояний на следующем шаге.

РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

Вычислительные эксперименты с целью выяснения зависимости затрат времени на моделирование нейросети от степени распараллеливания выполнялись на двух гетерогенных компьютерах. На первом было доступно 8 ядер процессора IntelCorei7-3930K 3.2GHz, графические процессоры NVIDIAQuadro 5000 (Q5000) и NVIDIAQuadroK5200 (Q5200). Второй имеет 4 ядра процессора IntelCore2 QuadQ8400 2.66GHz и графические процессоры AMDRadeonHD 5700 (R5700) и AMDRadeonHD 6900 (R6900).

Моделировалось три нейронные сети, содержащие один рабочий блок (9 слоев) и один управляющий блок (7 слоев). Основные количественные характеристики этих сетей сведены в таблицу 1.

Табл. 1. Характеристики нейросетей

	Нейронов	Синапсов	Память CPU (Мб)	Память GPU (Мб)	
Сеть 1	355820	46337146	242	236	232
Сеть 2	588820	80015815	418	408	404
Сеть 3	668820	90585206	473	462	456

Графики средних затрат времени на моделирование одного шага жизни сети (в секундах) при запуске модели:

–на одном ядре CPU без GPU,

–на всех ядрах CPU без GPU,

–на всех ядрах CPU и одном GPU (Q5000 или R5700),

–на всех ядрах CPU и одном GPU (Q5200 или R6900),

–на всех ядрах CPU и двух GPU,

приведены на рис.3 и рис.4.

Характер этих зависимостей, т.е. практически полное отсутствие эффекта ускорения вычислений за счет использования графических процессоров, может объясняться значительными затратами времени на пересылку больших (сотни мегабайт) массивов данных в/из памяти GPU на каждом шаге моделирования. Из памяти GPU в основную память поочередно пересылается один из массивов States0 и States1. Если симулятор после обработки полученного массива состояний нейронов модифицировал один или оба массива Synapses и Weights, то эти измененные параметры межнейронных связей переносятся из основной памяти в память каждого графического процессора.

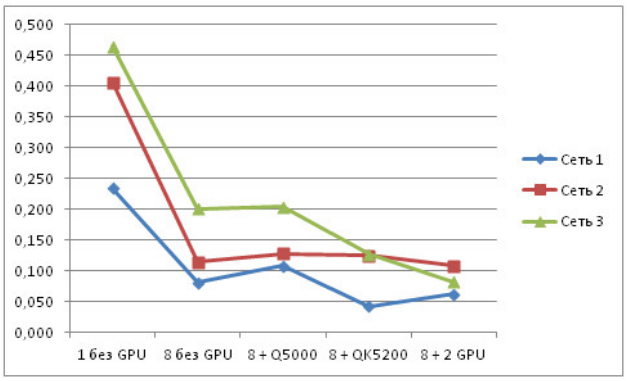

Рис. 3. Затраты времени первого компьютера

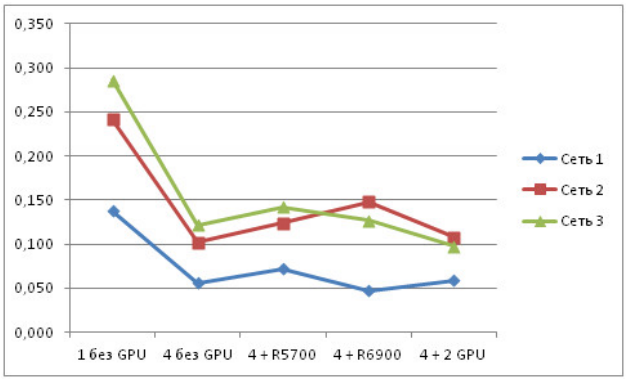

Рис. 4. Затраты времени второго компьютера

Для сети 1 существенное ускорение вычислений достигнуто только за счет использования нескольких ядер центрального процессора, подключение графических процессоров эффекта практически не дало. Но для сетей 2 и 3, содержащих большее количество нейронов, использование графических процессоров уже позволяет получить определенное, хотя и небольшое, ускорение.

ЗАКЛЮЧЕНИЕ

Рассмотрена задача моделирования многоблочной многослойной нейронной сети на гетерогенном компьютере с использованием всех или только некоторых составляющих его вычислительных устройств. С этой целью разработана программная модель, использующая технологии поддержки параллельных вычислений OpenMP и OpenCL. Показано, что для сетей с количеством нейронов до 400000 наибольшее ускорение может быть достигнуто при загрузке всех ядер центрального процессора и одного графического процессора. Для сетей с количеством нейронов порядка 500000 и более становится целесообразным привлекать к вычислениям уже несколько графических процессоров.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Anokhin P. Biology and neurophysiology of the conditioned reflex and its role in adaptive behavior, Ed. Oxford: Pergamon press, 1974

2. Addis M. New technologies and cultural consumption. Edutainment is born, Ed. Bocconi University: Marketing Department, 2002

3. Hawkins J. and Blakeslee S. On Intelligence, Ed. New York: Times Books, Henry Holt and Co., 2005.

4. Ahn, M., Lee, M., Choi, J. and Jun, S. C. A review of brain-computer interface games and an opinion survey from researchers, developers and users. Sensors 14(8), 2004

5. Maliavko A. and Gavrilov A. "Towards Development of Self-learning and Self-modification Spiking Neural Network as Model of Brain". *Actual problems of electronic instrument engineering (APEIE–2016)*, Novosibirsk, vol 1, part. 2, pp. 461-463, Oct. 2018.

6. Zhang D., Yao L., Zhang X., Wang S., Chen W. and Boots R. "Cascade and Parallel Convolutional Recurrent Neural Networks on EEG-Based Intention Recognition for Brain Computer Interface" *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*. pp 1703 – 1710, Apr. 2018.

7. Tsaregorodtsev V. "Parallel Implementation of Back-Propagation Neural Network Software on SMP Computers" *International Conference on Parallel Computing Technologies PaCT*, pp 186-192, Krasnoyarsk, Springer, 2005.

8. Martin E. and Cundy C. "Parallelizing linear recurrent neural nets over sequence length" Sixth International Conference on Learning Representations (ICLR), New Orleans, May 2018.

9. Minsky M. The emotion machine, Ed. New York : Simon&Shuster, 2006.

10. Gavrilov A. "Emotions and a prior knowledge representation in artificial general intelligence", *International Conference on Intelligent Information and Engineering Systems INFOS-2008*, ITHEA, Bulgaria. - pp. 106-110.

Малявко Александр Антонович – к.т.н., доцент кафедры вычислительной техники, Новосибирский государственный технический университет, тел. (383)3460492,e-mail: a.malyavko@corp.nstu.ru

PARALLEL IMPLEMENTATION OF A MULTILAYERED NEURAL NETWORK WITH ELEMENTS OF SELF-LEARNING ON A HETEROGENOUS COMPUTER

A.A. Maliavko

Novosibirsk state technical university, Novosibirsk

Abstract – The article discusses possible ways to reduce the time spent on the simulation of an artificial neural network, whose architecture is focused on the study of self-learning mechanisms. On the basis of comparison with biological neural systems that are known to be capable of self-learning, some assumptions are formulated about the possible structure of such a network as a combination of several functionally diverse types of multilayer neuron blocks. Connections between neurons are directed mainly from the network input to its output, but there are also connections between neurons of the same layer, as well as connections of the reverse direction. The effect of self-study may be achieved with the implementation of continuous cyclic modeling of the network, which is fully consistent with the mechanisms of functioning of biological prototypes. Continuous simulation of a network with a large number of neurons requires very large expenditures of computer time. Therefore, the focus is on the use of heterogeneous computers that provide significantly greater computing power compared to computers of traditional architecture. It describes a parallel software model developed for conducting experiments on the study of self-learning mechanisms on multi-core computers with several graphics processors, and the algorithm implemented in this model for the distribution and load balancing of graphics processors and cores of the central processor. The results of experiments on two different heterogeneous computers, showing a relatively weak effect of accelerating calculations if use of one or several GPUs, are presented. This effect can be explained by the need to constantly move large amounts of data between the main memory and the memory of graphics processors due to the continuous reconfiguration of the parameters of interneuron connections carried out by the simulator in the study of self-learning algorithms.

Index terms: neural network coordinates of the seat of fire, activation function, multipoint electro-optical system

REFERENCES

1. Anokhin P. Biology and neurophysiology of the conditioned reflex and its role in adaptive behavior, Ed. Oxford: Pergamon press, 1974
2. Addis M. New technologies and cultural consumption. Edutainment is born, Ed. Bocconi University: Marketing Department, 2002
3. Hawkins J. and Blakeslee S. On Intelligence, Ed. New York: Times Books, Henry Holt and Co., 2005.
4. Ahn, M., Lee, M., Choi, J. and Jun, S. C. A review of brain-computer interface games and an opinion survey from researchers, developers and users. Sensors 14(8), 2004
5. Maliavko A. and Gavrilov A. "Towards Development of Self-learning and Self-modification Spiking Neural Network as Model of Brain". *Actual problems of electronic instrument engineering (APEIE–2016)*, Novosibirsk, vol 1, part. 2, pp. 461-463, Oct. 2018.
6. Zhang D., Yao L., Zhang X., Wang S., Chen W. and Boots R. "Cascade and Parallel Convolutional Recurrent Neural Networks on EEG-Based Intention Recognition for Brain Computer Interface" *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18).* pp 1703 – 1710, Apr. 2018.
7. Tsaregorodtsev V. "ParallelImplementation of Back-Propagation Neural Network Software on SMP Computers" *International Conference on Parallel Computing Technologies PaCT*, pp 186-192, Krasnoyarsk, Springer, 2005.
8. Martin E. and Cundy C. "Parallelizing linear recurrent neural nets over sequence length" Sixth International Conference on Learning Representations (ICLR), New Orleans, May 2018.
9. Minsky M. The emotion machine, Ed. New York : Simon&Shuster, 2006.
10. Gavrilov A. "Emotions and a prior knowledge representation in artificial general intelligence", *International Conference on Intelligent Information and Engineering Systems INFOS-2008*, ITHEA, Bulgaria. - pp. 106-110.

Maliavko Aleksandr Antonovich – associate professor at the Dept. of Computer Engineering of the Novosibirsk State Technical University, Novosibirsk, (383)3460492, e-mail:a.malyavko@corp.nstu.ru