Применение нейронных сетей. II

А.А. ЯРОВОЙ
Винницкий национальный технический университет, Украина
axa@vinnitsa.com

ПРИКЛАДНЫЕ АСПЕКТЫ ПРОГРАММНО-АППАРАТНОЙ РЕАЛИЗАЦИИ НЕЙРОПОДОБНЫХ ПАРАЛЛЕЛЬНО-ИЕРАРХИЧЕСКИХ СИСТЕМ

Приводятся результаты исследований по разработке относительно новой вычислительной среды – параллельно-иерархической системы, которая предложена в виде сетевой модели нейроподобной схемы обработки информации. На основании исследуемого подхода формирования нейроподобной среды предложены математические модели, структурные и архитектурные решения, а также выбор программно-аппаратной базы для практической реализации нейроподобных параллельно-иерархических систем.

Методологические особенности организации нейроподобных параллельно-иерархических систем

Анализ последних работ по нейробиологии и работ, связанных с моделированием нейронных механизмов восприятия сенсорной информации показал, что остаются невыясненными следующие вопросы: каким образом происходит взаимодействие в коре головного мозга (ГМ) образующих нейроансамблей, их взаимодействие на уровне естественных локальных нейронных сетей; как во времени происходит интеграция пространственно разделенных активированных нейроансамблей Д. Хебба в горизонтальных и вертикальных путях в момент сочетающего действия многих одновременно действующих раздражителей? Поэтому в работе исследуется гипотетическая модель пространственной интеграции и структуризации информации в коре ГМ, относительно проблем в области обработки и распознавания образов [1]. Рассмотренные в работе модели пока что обладают в большей степени метафорическим сходством с „природными компьютерами", тем не менее, они предполагают новый более утонченный подход к машинным вычислениям, следуя которому можно будет создать новые микропроцессорные системы и компьютерные системы новой архитектуры. Кроме того, они позволяют по-новому взглянуть и на биологические системы. Прототипом предложенного подхода можно считать принципы коллективных вычислений в нейроподобных схемах кол-

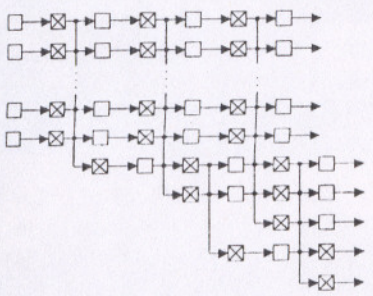
лективного принятия решений, которые требуют коллективного взаимодействия большого количества простых решений, в результате которого принимается сложное решение путем комбинирования данных на протяжении определенного промежутка времени. На основе анализа нейробиологических данных о теории структурирования сенсорной информации в мозге и особенностей организации вычислений в коре был выявлен ряд несоответствий относительно естественных механизмов восприятия объектов и ситуаций внешнего мира, которые не нашли соответствующего отображения в современных нейроподобных вычислителях, которые существенно ограничивают их технические возможности и не отвечают главным требованиям, которые относятся к интеллектуальным средствам обработки информации [2].

Исследуемый подход формирования нейроподобной параллельно-иерархической (ПИ) среды имеет ряд преимуществ по сравнению с другими методами формирования нейроподобной среды (например, по сравнению с известными методами формирования искусственных нейронных сетей). Главное преимущество подхода – это использование динамики многоуровневого параллельного взаимодействия информационных сигналов на разных уровнях иерархии нейроподобной сети, которая даёт возможность использовать такие известные естественные особенности организации вычислений в коре ГМ как: топографический характер отображения, одновременность (параллельность) действия сигналов, мозаичность структуры коры, грубую иерархичность коры, пространственно коррелированный во времени механизм восприятия и обучения [3].

Формирование многоэтапной ПИ сети предполагает процесс последовательного преобразования коррелированных и образования декоррелированных во времени элементов нейроподобной сети при переходе ее с одного устойчивого состояния в другое. Главной особенностью в предложенном подходе есть изучение динамики пространственно коррелированного механизма преобразования текущих и образования результирующих элементов нейронной сети. Такой механизм разрешает по-новому представить обработку в нейронной сети как процесс параллельно-последовательного преобразования различных составляющих изображения и учет временных характеристик преобразования. Причем, физическое содержание входных элементов нейронной сети, которые принимают участие в процессе корреляции-декорреляции, таких как, например, амплитуда или частота, фаза или энергия сигналов, связность или текстура изображений, определяется типом используемого преобразования, выбор которого зависит от класса решаемых задач. В общем виде концепцию

многоэтапности обработки изображений можно сформулировать так. Анализ изображения состоит в последовательном преобразовании совпадающих и обнаружении (фильтрации) несовпадающих во времени составляющих изображения при переходе элементов нейроподобной сети из текущих энергетических состояний с одними пространственными координатами в состояния с меньшей энергией с другими пространственными координатами. Такой процесс анализа изображения происходит на многих этапах, каждый из которых включает выполнение вышеуказанной процедуры. Условием перехода составляющих изображения на более высокий уровень является наличие динамики взаимного совпадения промежуточных результатов обработки во времени в параллельных каналах нижнего уровня. Результат анализа изображения формируется из изолированных в пространственно-временной области составляющих изображения [2,4].

Таким образом, исследуется нейроподобная сетевая структура (рис. 1), позволяющая имитировать принцип действия распределенной нейронной сети. Такая нейроподобная сетевая структура состоит из совокупности подсетей (рис. 2) формирования признаков состояния пространственно-временной среды (ПВС), структура которых однородна и состоит из ряда взаимозависимых иерархических уровней [3].



☐ – множество различных состояний ПВС;
☒ – общие состояния ПВС.

Рис. 1. Структура параллельно - иерархической сети

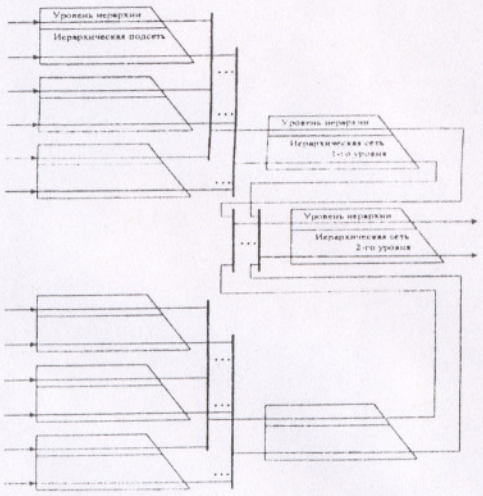


Рис. 2. Структура многоступенчатой иерархической сети

Алгоритм работы сети универсальный и состоит в параллельно-иерархическом формировании совокупностей общих и разнообразных признаков-сигналов состояния ПВС. Обобщения всех видов сенсорной информации происходит на самом конечном этапе преобразования вне иерархической обработки каждого вида сенсорной информации. "Интеллектуальный" уровень распределенной сети определяется степенью обобщения сенсорной информации в ее ветвях. Чем больше степень обобщения сенсорной информации при ее прохождении по ветвям сети, тем выше ее "интеллектуальный" уровень. В каждой ветви ПИ сети реализуется алгоритм пирамидальной обработки. Сущность ПИ подхода состоит в одновременном использовании последовательности множеств массивов данных, которые образовывают множества информационных полей на различных уровнях иерархии, рекурсивном формировании новых последовательностей информационных потоков на различных уровнях иерархии, что разрешает реализовать стратегию многоуровневого взаимодействия от "общего к частному". В общей форме методологию формирования ПИ сети можно представить в формализованном виде, исходя из следующих положений [3]. Пусть дано множество потоков входных данных. Возникает следующая проблема. Каким образом так организовать параллельный вычислительный процесс, чтобы получить строго распределенную в времени и иерархии нейроподобную сеть? Если обрабатывать множество входных потоков данных на разнообразных *(k)* иерархических уровнях, то каждый уровень представляет собой совокупность процессорных элементов, которые функционируют в строго фиксированные моменты времени (t_j). Пусть заданны n_1 функций $f_1(t), f_2(t),..., f_{n_1}(t)$. Данные функции опишем на разных уровнях иерархии от 1-го до j-го $(j = 2l, l = 1,2...$ и $j = 2l+3,\ l = 0,\ 1,\ 2,...)$.

$$\sum_{j=1}^{n_1}\sum_{i=1}^{n_{j1}} f_{j1}\left(t-i\tau\right) = \sum_{j=1}^{n_2}\sum_{i=1}^{n_{j2}} f_{j2}\left(t-2ij\tau\right) = \sum_{i=1}^{n_{1(2l+3)}} f_{1(2l+3)}\left(t-\left(2i+6l+3\right)\tau\right) +$$

$$+ \sum_{i=2}^{n_{2(2l+3)}} f_{2(2l+3)}\left(t-\left(2i+6l+3\right)\tau\right) + \ldots + \sum_{i=n-1}^{n_{(n-1)(2l+3)}} f_{(n-1)(2l+3)}\left(t-\left(2i+6l+3\right)\tau\right) +$$

$$+ \sum_{i=n}^{n_{n(2l+3)}} f_{n(2l+3)}\left(t-\left(2i+6l+3\right)\tau\right) = \sum_{j=2}^{k} f_{i,j}\left\{t-(3j-4)\tau\right\}, \qquad (1)$$

где τ – задержка формирования следующей функции относительно предыдущей, n_{jk} – число функций j-го разложения k-го функционального уровня. Правая часть выражения (1) формирует хвостовые функции, которые во времени распределены по разным иерархическими уровням и были получены в результате этого функционального преобразования. Анализируя сетевое преобразование (1), можно сделать вывод о том, что в процессе образования каждого уровня формируется временной сдвиг (τ), наличие которого приводит к получению хвостовых функций.

Анализ подходов для программно-аппаратной реализации нейроподобных параллельно-иерархических систем

В проведенных исследованиях был проведен сравнительный анализ наиболее распространенных технологий в качестве аппаратных платформ реализации нейроподобных параллельно-иерархических систем, а именно центральные процессоры (CPU), нейрочипы и видеоадаптеры (GPGPU), поскольку их использование разрешает абстрагироваться от уровня проектирования аппаратной платформы. Кроме рассмотренных средств, существуют также другие пути решения поставленной задачи – например, собственноручное изготовление схемы на основе DSP-процессоров, но данное решение имеет весомый в нашем случае недостаток – привязанность к определенной топологии [5]. На основании проведенных исследований, было принято решение выбрать в качестве аппаратной платформы для реализации нейроподобных параллельно-иерархических систем видеоадаптеры (GPGPU) [6].

В целом, использование видеоадаптера для вычислений общего назначения мало чем отличается от эмуляции на центральном процессоре. Тем не менее, есть существенная разность – программа, которая использует видеоадаптер, для максимальной эффективности (утилизации аппаратных ресурсов) должна быть параллельная относительно данных или задач (так называемые „Data Parallelism" и „Task Parallelism"). При этом основной блок вычислений программы компилируется в байт-код DirectX 9 или 10, или в соответствующий байт-код ATI CTM IL. Такой байт-код транслируется в специальный машинный код (так называемый „device-specific assembler") перед выполнением. Рассмотрим аппаратную базу: современные массовые видеоадаптеры по своему теоретическому быстродействию превышают современные процессоры в 10-20 раз, количество загрузок из памяти значительно больше, что объясняется большей шириной шины и более высокой тактовой частотой памяти. Видеоадаптеры, в отличие от нейрочипов, являются массовым продуктом (больше того – продуктом

большого спроса), а потому они изготовляются по актуальному техническому процессу и широко доступны [7].

Конечно, характерным недостатком такого подхода является сложность реализации алгоритма определенной прикладной задачи в параллельной относительно данных форме – соответственно сложно достичь уровня теоретического быстродействия. Кроме того, быстродействие вычислений ограничивается эффективностью обращений к памяти (основные факторы: пропускная способность локальной памяти и задержки при обращении к основной памяти). Для некоторых задач может играть важную роль малый объем локальной памяти (до 4 Гб, и 512 Мб в среднем). Но, несмотря на приведенные недостатки, технологии GPGPU избраны в исследовании для аппаратной реализации нейроподобных параллельно-иерархических систем [6]. Рассмотрим конкурирующие решения [7].

Продукция компании NVidia

Американская компания NVidia предлагает несколько вариантов решений для выполнения вычислений общего назначения, которые базируются на технологии CUDA. CUDA (англ. Compute Unified Device Architecture) – технология GPGPU, которая разрешает программистам реализовывать на языке программирования Си алгоритмы, выполняемые на графических процессорах ускорителей GeForce восьмого поколения и старше (GeForce 8 Series, GeForce 9 Series, GeForce 200 Series), Nvidia Quadro и Tesla компании NVidia.

Рассмотрим ниже показатели мощнейшей видеокарты – NVidia GeForce GTX280 1 GB в сравнении с соответствующей видеокартой AMD/ATI -ATI Radeon HD 4870 1GB.

Продукция компании ATI

Канадская компания ATI Technologies (с 2007 года подраздел AMD Graphics Products Group) – один из наибольших производителей и поставщиков графических процессоров в мире. В 2006 году ATI была приобретенная корпорацией AMD, тем не менее, бренд ATI все еще сохраняется на рынке видеоадаптеров.

Для вычислений общего назначения, как и компания NVidia, ATI выпускает серию видеокарт-вычислителей (ATI FireStream) и допускает выполнение таких вычислений на игровых видеокартах, начиная с серии R580 (ATi Radeon X1900 XTX).

Сравним некоторые из лидирующих на данный момент на рынке видеокарты обозначенных компаний по следующим критериям (табл. 1).

Учитывая удельную стоимость быстродействия, видеоадаптеры ATI являются наиболее оптимальным решением для вычислений общего характера и соответственно для реализации нейроподобных параллельно-иерархических систем.

Таблица 1

Критерий	NVidia	ATI
Максимальное теоретическое быстродействие	1 Tflops	1,2 Tflops
Пропускная способность памяти	141,7 GB/s	115,2 GB/s
Цена	520 USD*	320 USD*
Удельное быстродействие	1,92 Gflops/USD	3,75 Gflops/USD
Удельная пропускная способность памяти	0,27 GB/s/USD	0,36 GB/s/USD

* Средняя цена взята по данным сайта www.hotline.ua от 12.10.2008.

Анализ программных платформ для GPGPU

В качестве программных платформ для реализации вычислений общего характера на видеоадаптерах можно выделить такие широко используемые варианты [7]:

- Ассемблер (ATI CTM IL);
- Шейдерные языки (GLSL-OpenGL 2.0, HLSL-DirectX 9.0c+);
- Высокоуровневые языки (NVidia CUDA, RapidMind, Brook/Brook+).

Сравнительная характеристика программных платформ GPGPU представлена в табл. 2 [6].

В результате приведенного выше анализа, была поставлена первичная задача – создания программной библиотеки, которая бы разрешала обрабатывать как масштабные искусственные нейронные сети, так и нейроподобные параллельно-иерархические сети. Первым шагом, который уже осуществлен, стало создание библиотеки, которая разрешает обрабатывать искусственной нейронные сети произвольной топологии, а также нейроподобные параллельно-иерархические сети на центральном процессоре (CPU). Следующий шаг – создания GPU-версии данной библиотеки. На данный момент эта версия находится в стадии тестирования и отладки [6].

Особенностью такой реализации есть то, что алгоритм обработки передачи импульса между тактами с учетом специфики программирования параллельных устройств (чтобы избежать реализации механизма синхро-

низации параллельных потоков) нуждается в преобразовании формата данных, которое происходит таким образом [6]:

Таблица 2

Возможности	ATI CTM IL	GLSL/ HLSL	NVidia CUDA	Rapid Mind	Brook/ Brook+
Произвольное считывание из памяти	+	+	+	+	+
Произвольная запись в память	+	–	+	+	– / +
Разрядность	64 bit	32 bit	64 bit (CUDA 2.0)	32 bit	64 bit
Лицензия	Freeware	Freeware	Freeware	Shareware (демоверсия отсутствует)	Open source
Поддержка видеоадаптеров	ATI (2XXX+)	Любой OpenGL 2.0-совместный (GLSL) / любой DirectX 9.0c-совместный (HLSL)	NVidia (8XXX+)	Любой DirectX 10-совместный	Любой DirectX 9.0c или OpenGL 2.0 совместный (Brook) / ATI (серия 2XXX+) (Brook+)
Возможность низкоуровневой оптимизации	+	–	–	–	– / +
Не нуждается в среде выполнения	+	–	+	–	–

1. Для каждого нейронного элемента строится одноизмеримая таблица, каждый элемент которой является структурой „номер связанного нейрона в предыдущем слое – вес межнейронной связи".

2. Таким образом, для каждого слоя текущего такта получается набор таблиц по количеству нейронов в шаре, которые характеризуют межнейронные связи.

3. Кроме того, для каждого слоя строится дополнительная одноизмеримая таблица, которая включает в себя уровни активации нейронных элементов данного слоя.

Такое преобразование разрешает сохранять данные, необходимые для передачи импульса между тактами, в едином массиве и загружать их в память видеокарты за один цикл передачи данных. Такая реализация разрешает избежать использования операции произвольной записи в память, которая не поддерживается видеокартами младше серии R670, и реализации механизма синхронизации между параллельными потоками.

Выводы

Таким образом, в результате проведенных исследований был проведен сравнительный анализ программно-аппаратных вариантов реализации нейроподобных параллельно-иерархических систем, которые по уровню сложности не уступают крупномасштабным искусственным нейронным сетям. Выбран, на основании критериев быстродействия и стоимости, наиболее подходящий вариант для эффективной прикладной реализации не только нейроподобных параллельно-иерархических систем, но и классических парадигм искусственных нейронных сетей. Это подтверждается полученными результатами программных экспериментов, полученных в результате моделирования задач прогнозирования, распознавания изображений на разработанных оригинальных программных продуктах под такие аппаратные платформы как CPU и GPGPU [8-10].

Список литературы

1. В.П. Кожем'яко, Л.І. Тимченко, А.А. Яровий Паралельно-ієрархічні мережі як структурно-функціональний базис для побудови спеціалізованих моделей образного комп'ютера. Монографія. – Вінниця: Універсум-Вінниця, 2005. – 161 с.

2. Кожем'яко В.П., Лисенко Г.Л., Яровий А.А., Кожем'яко А.В. Образний відеокомп'ютер око-процесорного типу. Монографія. – Вінниця: Універсум-Вінниця, 2008. – 215 с.

3. Паралельно-ієрархічне перетворення як системна модель оптико-електронних засобів штучного інтелекту. Монографія / В.П. Кожем'яко, Ю.Ф. Кутаєв, С.В. Свєчніков, Л.І. Тимченко, А.А. Яровий – Вінниця: УНІВЕРСУМ-Вінниця, 2003. – 324 с.

4. М.В. Ковзель, Л.І. Тимченко, Ю.Ф. Кутаєв, С.В. Свєчніков, В.П. Кожем'яко, О.І. Стасюк, С.М. Білан, Л.В. Загоруйко Паралельно-ієрархічне перетворення і Q-обробка інформації для систем реального часу. Монографія. Кол. авторів, - Київ: "КУЕТТ", 2006. – 492 с.

5. Кожем'яко В.П., Тимченко Л.І., Яровий А.А., Ремезюк С. Апаратна реалізація паралельно-ієрархічної мережі на основі DSP – Оптоелектронні інформаційні технології "Фотоніка ОДС–2005". Збірник тез доповідей третьої міжнародної науково-технічної конференції, м. Вінниця, 27-28 квітня 2005 року. – Вінниця: „УНІВЕРСУМ-Вінниця", 2005. – С. 43.

6. А.А. Яровий, Ю.С. Богомолов, К.Ю. Вознесенський. Вибір апаратної платформи для реалізації масштабних нейронних та нейроподібних паралельно-ієрархічних мереж - IX Міжнародна конференція Контроль і управління в складних системах (КУСС-2008), Вінниця, 21-24 жовтня 2008 року – Mode of access: World Wide Web http://www.vstu.vinnica.ua/mccs2008/ /materials/subsection_2.2.pdf.

7. GPGPU: General Purpose Computing Using Graphics Hardware – Mode of access: World Wide Web http://www.gpgpu.org.

8. Kozhemyako V.P., Timchenko L.I., Yarovyy A.A. Parallel-Hierarchical Transformation as the System Model of Neurolike Scheme of Data Processing. Сборник трудов IV Международной конференции «Фундаментальные проблемы оптики-2006». Санкт-Петербург. 16–20 октября 2006 / Под ред. проф. В.Г. Беспалова, проф. С.А. Козлова – СПб.: Издательский дом «Corvus», 2006. – с. 246-248.

9. V.P. Kozhemyako, L.I. Timchenko, A.A. Yarovyy Methodological Peculiarities of Neural-Like Network Model for Pyramidal and Parallel-Hierarchical Processing of Digital Information. Proceedings of the 9th International Conference on „Development and application systems (DAS-08)", May 22-24, 2008, Suceava, Romania – Suceava, Universitatea Stefan cel Mare Suceava, 2008. – p. 313-319.

10. L.I. Timchenko, Yu.F. Kutaev, V.P. Kozhemyako, A.A. Yarovyy, A.A. Gertsiy, A.T. Terenchuk, Nafez Shweiki Method for Training of a Parallel-Hierarchical Network, Based on Population Coding for Processing of Extended Laser Paths Images – USA, Proceedings of SPIE, Volume 4790, 2002. – p. 465-479.

Д.К. КУЙКИН, В.В. ХРЯЩЕВ

Ярославский государственный университет им. П.Г. Демидова

denis.kuykin@gmail.com, vladimir@piclab.ru

НЕЙРОННАЯ СЕТЬ В ЗАДАЧЕ НЕЭТАЛОННОЙ ОЦЕНКИ КАЧЕСТВА СЖАТЫХ ИЗОБРАЖЕНИЙ

Описывается нейросетевая модификация алгоритма неэталонной оценки качества сжатых изображений формата JPEG. Приведенные результаты моделирования показывают хорошую коррелированность получаемых оценок с визуально воспринимаемым качеством декодированных изображений.