УДК 004.932

В. А. КОЛБАСИН, канд. тех. наук, доц., НТУ «ХПИ»;
И. А. ХРИСТЕНКО, магистрант, НТУ «ХПИ»;
Д. ***А. ХРИСТЕНКО***, магистрант, НТУ «ХПИ»

ПАРАЛЛЕЛЬНАЯ РЕАЛИЗАЦИЯ АЛГОРИТМА ОПТИЧЕСКОГО РАСПОЗНАВАНИЯ РУКОПИСНЫХ СИМВОЛОВ НА ПЛАТФОРМЕ CUDA

У статті розглядаються питання організація паралельних обчислень на платформі CUDA для задачі оптичного розпізнавання рукописних символів з використанням штучних нейронних мереж (ШНМ). У роботі використовується ШНМ прямого поширення з двома прихованими шарами. Розглянуто два способи організації паралельних обчислень: один потоковий процесор обробляє дані одного символу, і все потокові процесори карти обробляють дані одного символу. Показано, що перший підхід до розпаралелювання є більш ефективним.

В статье рассматриваются вопросы организация параллельных вычислений на платформе CUDA для задачи оптического распознавания рукописных символов с использованием искусственных нейронных сетей (ИНС). В работе используется ИНС прямого распространения с двумя скрытыми слоями. Рассмотрено два способа организации параллельных вычислений: один потоковый процессор обрабатывает данные одного символа, и все потоковые процессоры карты обрабатывают данные одного символа. Показано, что первый подход к распараллеливанию является более эффективным.

The article focuses on organization of parallel computing using CUDA platform for optical character recognition problem of handwritten characters using artificial neural networks (ANN). The backpropagation ANN with two hidden layers was used. Two ways of organizing parallel computing were considered: a stream processor handles data of one character, and all the stream processors handles the data of one character. It is shown that the first approach to parallelization is more efficient.

Введение. В настоящее время системы оптического распознавания символов (OCR) являются незаменимым инструментом офисных работников, имеющих дело с большим объёмом документов. Они широко используются для автоматизированного приведения бумажных документов в форму, пригодную к компьютерной обработке содержащегося в них текста.

Системы OCR обеспечивают многократное увеличение скорости ввода документов по сравнению с перепечаткой документа вручную. Но, такое ускорение обеспечивается только при вводе печатного текста, так как распознавание рукописных символов большинством коммерческих систем OCR не поддерживается.

Одной из причин такого положения дел является то, что изменчивость начертания букв рукописного текста вынуждает использовать для его распознавания намного более сложные и ресурсоемкие методы. Соответственно увеличивается время обработки документа и снижается практическая ценность OCR-системы, т.к. время подготовки текстового представления оказывается сравнимым со временем перепечатывания текста вручную. Поэтому

уменьшение времени обработки исходного графического документа системой OCR является важной практической задачей.

Искусственные нейронные сети (ИНС) широко применяются для решения задач распознавания образов и в частности, для распознавания рукописных символов [1, 2]. Применение ИНС позволяет добиться приемлемого качества распознавания, но требует значительных затрат вычислительных ресурсов. Одним из способов решения данной проблемы является использование технологий параллельных и распределенных вычислений, таких как кластерные системы, аппаратные нейропроцессоры, использование процессоров видеокарт для неграфических вычислений и т.д.

Так как основная сфера применения систем OCR – это офис, где применение дорогостоящих и специализированных решений неприемлемо по экономическим соображениям, наибольший практический интерес представляет использование процессоров видеокарт для выполнения распознавания. Наиболее документированной и удобной для разработчика является технология массовых параллельных вычислений на процессорах видеокарт CUDA [3]. С ее помощью удается достичь высокой производительности системы при относительно малой стоимости решения для задач, которые могут быть представлены в модели SIMD (одна инструкция выполняется над многими данными).

На данный момент существует достаточно большое число исследований, посвященных созданию параллельных реализаций методов распознавания одного рукописного символа для разных вычислительных архитектур, но мало исследований, посвященных проблеме организации параллельных вычислений для распознавания всего набора символов документа. Так как последовательное применение параллельных алгоритмов распознавания для каждого символа документа не является наилучшим способом организации вычислений, решение данной задачи представляет существенный практический интерес. Поэтому данная работа посвящена разработке параллельной реализации алгоритма распознавания набора рукописных символов ИНС для платформы CUDA и исследованию влияния подхода к организации параллельных вычислений на скорость обработки данных. В качестве тестового набора рукописных символов используется общепринятая библиотека изображений рукописных цифр MNIST [4].

Распознавание изображений с помощью искусственной нейронной сети. Для распознавания символов в данной работе используется многослойный перцептрон, структура которого представлена на рисунке 1. Входной слой содержит 784 нейрона для представления данных изображения размером 28×28 пикселей, два скрытых слоя содержат 1000 и 500 нейронов соответственно. Число нейронов выходного слоя соответствует числу распознаваемых символов и равняется 10.

На вход сети подается нормированное изображение рукописного символа размером 28×28 пикселей.

Значение на выходе каждого нейрона скрытых и выходных слоев вычисляется по формуле [2]:

$$y_i^{(l)} = f\left(\sum_{j=0}^{N-1} x_j^{(l)} \cdot w_{i,j}^{(l)}\right), \quad (1)$$

где l – номер слоя;

$x_j^{(l)}$ – значение на j-м входе i-го нейрона слоя l;

$w_{i,j}^{(l)}$ – весовой коэффициент;

$y_i^{(l)}$ – значение на выходе i-го нейрона;

$f(z) = 1/\left(1+e^{-x}\right)$ – функция активации.

Данный перцептрон был обучен с помощью стандартного метода обратного распространения ошибки [1]. После обучения перцептрона сеть стала распознавать 97.9% символов из тестовой выборки.

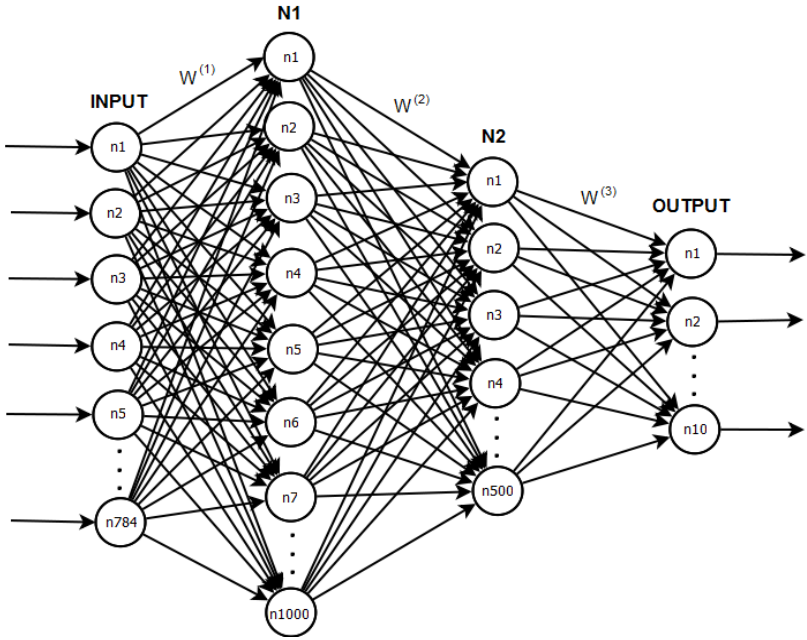


Рис. 1 – Структура используемого перцептрона

Оптимизация вычислений с использованием CUDA. В данной работе рассматривается только обработка данных ИНС в процессе их распознавания. Ускорение процесса обучения сети здесь не рассматривается, так как обучение выполняется только один раз, при настройке сети на набор рукописных символов, а обработка выполняется каждый раз при распознавании нового документа.

Процессоры видеокарт, поддерживающих NVidia CUDA, содержат большое число параллельных узлов – потоковых мультипроцессоров (SMP), каждый из которых работает как SIMD-вычислитель. Программно работа с вычислителем CUDA представлена блочно-сеточной моделью организации параллельных вычислений, при которой все нити, выполняющие единую функцию-ядро (Kernel), объединяются в блоки (Block), а блоки, в свою очередь, объединяются в сетку (Grid). Потоки, объединенные в один блок, выполняются на одном SMP, поэтому для увеличения производительности желательно использовать сетку с числом блоков, кратным количеству SMP в устройстве. Если для выполнения блока требуется малое количество регистров и разделяемой памяти, то на одном SMP могут выполняться несколько блоков. Память платформы CUDA разделена на несколько частей: глобальная память компьютера, память платформы (видеокарты), кэш доступа к памяти и так называемая разделяемая память (shared memory). Виды памяти перечислены в порядке уменьшения времени доступа к ней. При осуществлении доступа к областям памяти, в которых хранятся константы и текстуры осуществляется кэширование. Объем кэша зависит от версии платформы CUDA, но не превышает 8Кб на один SMP. Разделяемая память размещается на SMP и обеспечивает минимальное время доступа, но имеет объем всего в 16 Кб и разделена на банки. Наибольшая производительность при работе с разделяемой памятью достигается, когда каждый поток обращается к своему банку памяти.

В данной работе было использовано два подхода к распараллеливанию распознавания символов ИНС.

Первый подход предполагает организацию вычислений по схеме «одно изображение – один потоковый процессор». Для этого обработка данных одного изображения осуществляется в одном блоке потоков. Каждый поток вычисляет значения на выходе одного или нескольких нейронов. Число блоков в ядре определяется исходя из требований к скорости реакции системы и типового объема обрабатываемых данных. Использование большего числа блоков, чем количество установленных на аппаратуре потоковых мультипроцессорах (SMP), позволяет уменьшить затраты на запуск потоков и, если достаточно ресурсов, запустить несколько блоков выполняться на одном SMP. Для эффективного использования ресурсов карты число потоков должно равняться $k \cdot N$, где k – целое число, а N – число SMP в используемой видеокарте.

При вычислении значения выходов нейронов по формуле (1) весовые коэффициенты входов нейронов будут участвовать в обработке одного окна данных только один раз. Значения, подаваемые на вход слоя ИНС, напротив, будут использоваться много раз всеми нейронами этого слоя, поэтому для их хранения имеет смысл использовать быструю разделяемую память. Таким образом, в разделяемой памяти размещаются значения входов и выходов каждого слоя нейронов. Весовые коэффициенты входов нейронов располагаются в глобальной памяти, а доступ к ним осуществляется через кеш механизма чтения текстур.

Объем разделяемой памяти позволяет разместить в ней незначительную часть синаптических весов нейронов – 2200 значений. Для проверки того, как повлияет такое решение на быстродействие реализации, была создана дополнительная реализация первого подхода к распараллеливанию.

Второй подход предполагает организацию вычислений по схеме «одно изображение – вся карта CUDA». При такой организации вычислений потоки каждого SMP будут обрабатывать свою часть данных. В этом случае как входные, так и промежуточные данные должны храниться в глобальной памяти устройства. Также здесь возникает проблема синхронизации потоковых процессоров между собой. Дело в том, что встроенные средства CUDA позволяют синхронизировать только выполнение потоков внутри одного блока потоков (все потоки которого выполняются на одном SMP). Для синхронизации потоков, принадлежащим разным блокам встроенных средств нет. А при обработке данных ИНС требуется синхронизировать вычисление значений на выходах каждого слоя сети: вычисление значений нейронов следующего слоя должно начаться только после того, как закончится вычисление значений нейронов предыдущего слоя. Для решения этой проблемы в работе используется не самый быстрый, но самый простой способ: вычисление значений нейронов каждого слоя выполняется отдельным ядром (kernel).

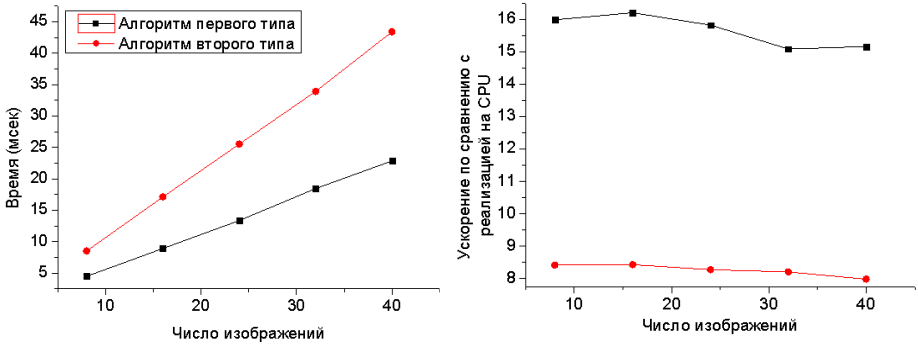
При использовании второго подхода разделяемая память используется для хранения входных значений слоя и для хранения части весовых коэффициентов.

Результаты. Скорость работы предложенных алгоритмов исследовалась с использованием поддерживающей технологию CUDA видеокарты GeForce 9600GT, которая содержит 8 потоковых мультипроцессоров (SMP).

Для определения скорости работы предложенных методов измерялось время обработки наборов тестовых изображений. Так как карта имеет 8 SMP, число изображений в наборе также было кратным 8. Для алгоритмов первого и второго типа сравнивалось время распознавания набора одинаковых изображений. Результаты сравнения приведены на графиках на рисунке 2.

Как видно из приведенных графиков, за счет больших затрат ресурсов на запуск ядер и менее эффективного управления памятью алгоритмы второго типа практически в два раза медленнее алгоритмов первого типа. Однако обе

CUDA реализации обрабатывают данные почти в восемь раз быстрее, чем реализация на центральном процессоре Intel Celeron G530 2.4 GHz.



а *б*

Рис. 2 – Зависимость времени обработки набора изображений от типа алгоритма (*а*); и ускорение по сравнению с реализацией на CPU (*б*)

В целом, предложенная параллельная реализация алгоритма распознавания рукописных символов при помощи многослойной ИНС прямого распространения позволяет выполнить распознавание символов намного быстрее, чем реализация на центральном процессоре, и может быть использована при разработке систем оптического распознавания символов.

Список литературы: **1.** *Бодянский Е. В.* Искусственные нейронные сети / *Е. В. Бодянский, О. Г. Руденко*. – Х. : Компания СМИТ, 2005. – 408 с. **2.** *Осовский С.* Нейронные сети для обработки информации / *С. Осовский*. – М. : Финансы и статистика, 2004. – 344 с. **3.** *Боресков А. В.* Основы работы с технологией CUDA / *А.В. Боресков, А. А. Харламов*. – М. : ДМК Пресс, 2010. – 232 с. **4.** *LeCun Y.* The MNIST database of handwritten digits [Электронный ресурс] // *Y. LeCun, C. Cortes*. – 2009. – Режим доступа: http://yann.lecun.com/exdb/mnist/index.html.

Надійшла до редколегії 08.05.2012