NVIDIA

CUDA® АЛЬМАНАХ

АВГУСТ 2015

СОДЕРЖАНИЕ

ПРОГНОЗИРОВАНИЕ ЭПИДЕМИЙ С ИСПОЛЬЗОВАНИЕМ GPU

Крис Джуэлл, первоначально получивший образование ветеринарного хирурга, а затем преподававший эпидемиологию в Медицинской школе Ланкастера (Великобритания), еще в 2001 году заинтересовался этой темой во время работы над эпидемией ящура в 2011 году. Его деятельность была связана с изучением таких заболеваний животных, как ящур, тейлериоз и птичий грипп. Исследования проводились совместно с государственными организациями Великобритании, Новой Зеландии, Австралии и США. Недавно Крис начал заниматься эпидемиями, распространяющимися среди людей. Поэтому ему понадобилось больше вычислительных ресурсов.

В настоящее время Крис использует технологию CUDA для ускорения расчетов по прогнозированию эпидемий в режиме реального времени.

«Без технологии CUDA метод Монте-Карло с цепями Маркова, который мы используем, работает слишком медленно для практичного использования. Мы получили ускорение в 380 раз по сравнению с одноядерным CPU кодом. Это значит, что прогнозирование в реальном времени теперь стало возможным». Подробнее.

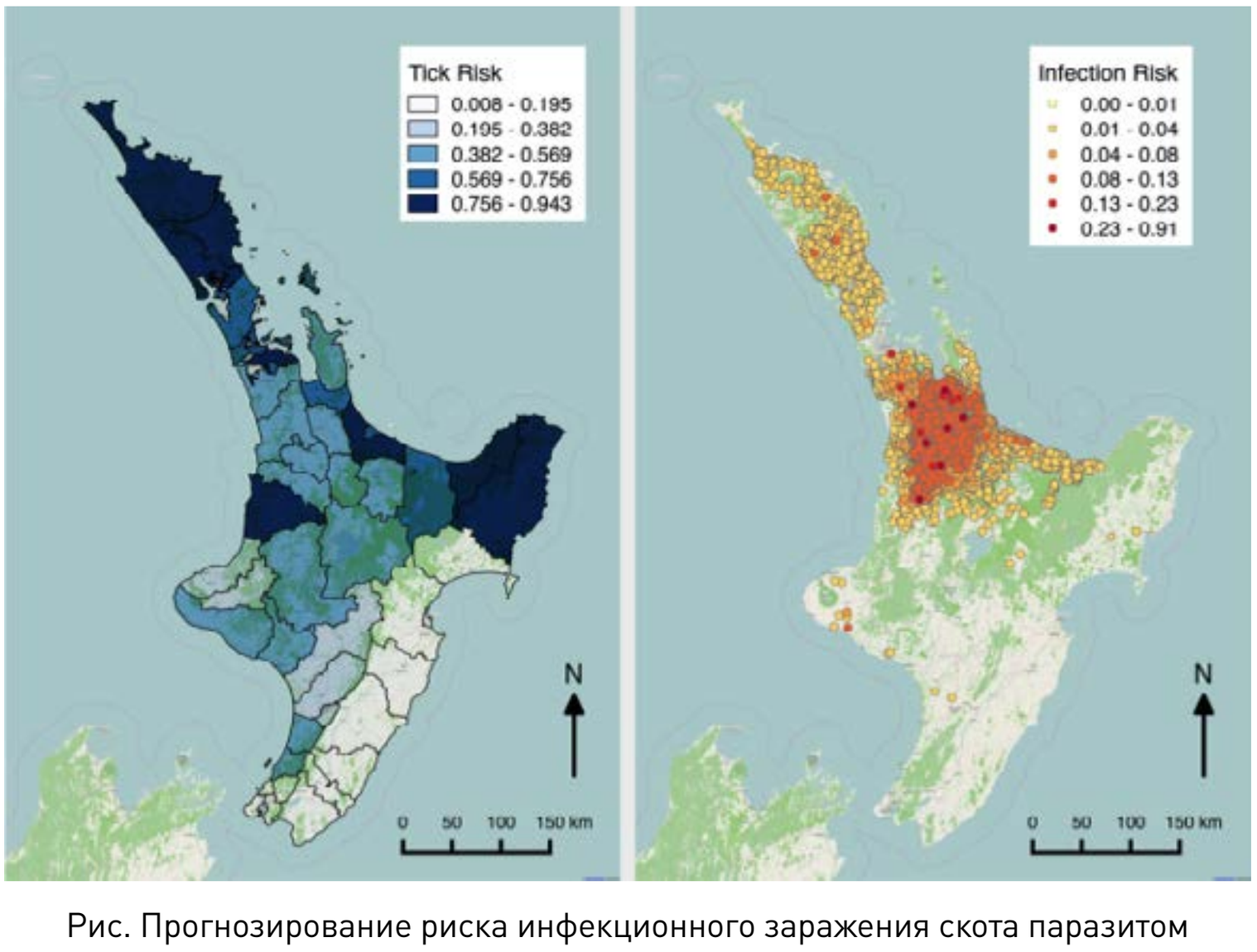
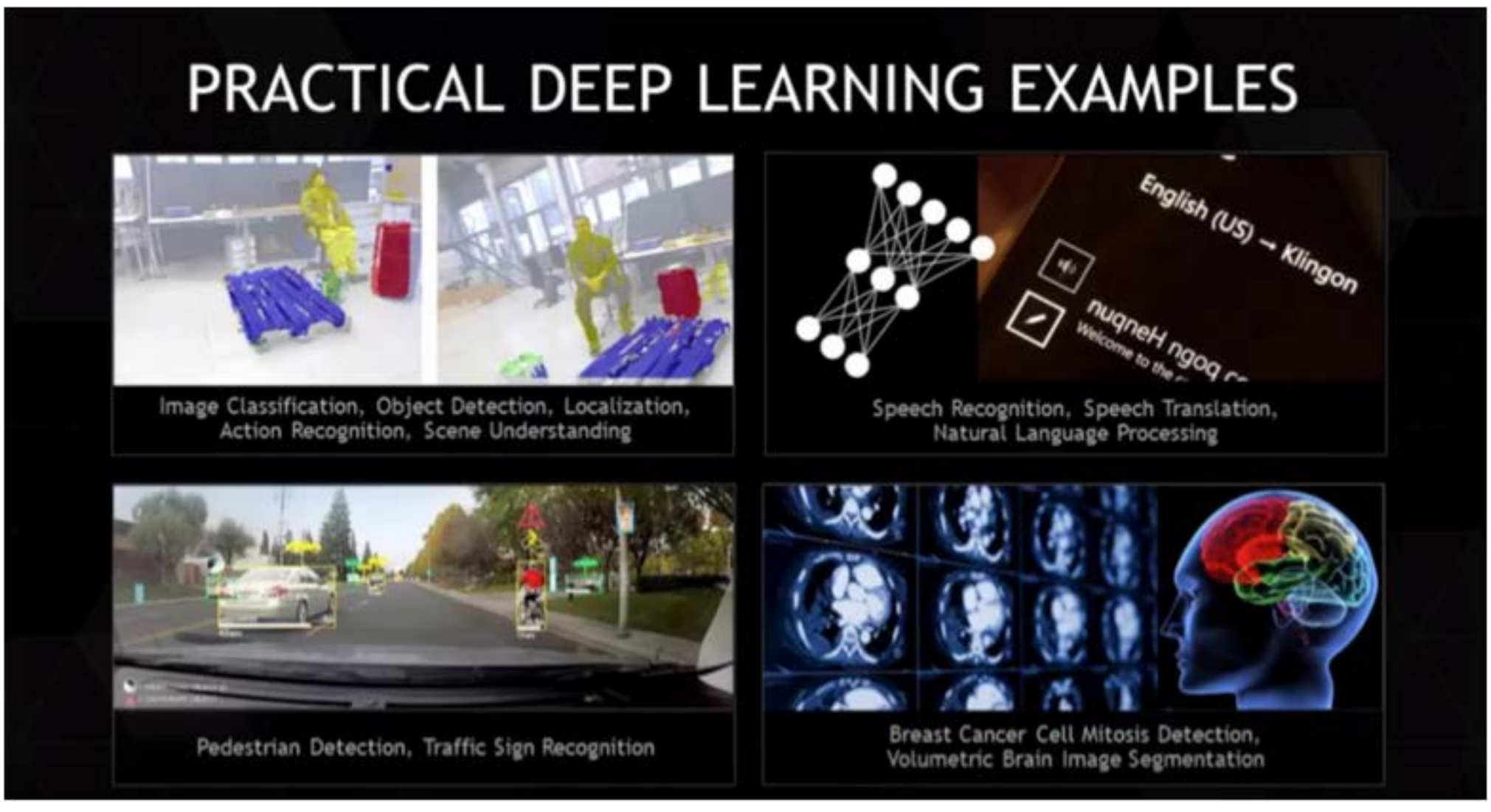


Рис. Прогнозирование риска инфекционного заражения скота паразитом Theileria orientalis среди неифицированных животных

БЕСПЛАТНЫЙ ОНЛАЙН КУРС ПО ГЛУБОКОМУ ОБУЧЕНИЮ

У вас есть прекрасная возможность пройти курс по глубокому обучению от NVIDIA. На сайте доступны записи лекций, презентации и практические задания.



Подробнее о машинном обучении можно узнать здесь.

СУПЕРКОМПЬЮТЕРНЫЕ ДНИ В РОССИИ

Суперкомпьютерный консорциум университетов России и Федеральное агентство научных организаций России 28-29 сентября 2015 г. проводят в Москве международную конференцию.

Кнференция рассчитана на самый широкий круг представителей науки, промышленности, бизнеса, образования, государственных органов, учащихся — всех тех, кто связан с разработкой или использованием суперкомпьютерных технологий. Тематика конференции охватывает все аспекты суперкомпьютерных технологий: разработка аппаратного и программного обеспечения, решение больших задач, использование суперкомпьютерных технологий в промышленности, проблемы экзафлопсных вычислений, суперкомпьютерное образование и многие другие.

В первый день работы конференции будет объявлена 23-я редакция списка Top50 самых мощных компьютеров СНГ.

У вас будет возможность узнать про суперкомпьютерные решения NVIDIA, пообщаться с представителями компании, а также пройти семинар по программированию GPU для начинающих.

Подробнее

ВЕБИНАРЫ НА АНГЛИЙСКОМ ЯЗЫКЕ

8 Сентября: More Science, Less Programming with OpenACC

8 Октября: Applied Deep Learning for Vision and Natural Language with Torch7

Также доступны для просмотра:

Ускорение вычислений с OpenACC

CUDA 7 - обзор производительности

CUDA 7 – описание и особенности

Ускорение глубокого обучения с помощью cuDNN

Обзор DIGITS

Основные моменты GTC 2015

РАСПОЗНАВАНИЕ СЦЕН И ДОМАШНИХ ЖИВОТНЫХ НА ИЗОБРАЖЕНИЯХ В ДЕСКТОПНОМ ФОТООРГАНАЙЗЕРЕ ZZ PHOTO НА ОСНОВЕ DEEP LEARNING

Ю.А. Пащенко, А.Н. Чернодуб

Подробнее

Глубокое обучение (Deep Learning) – лидирующая сейчас технология в области распознавания изображений. Глубокие сверточные нейронные сети используют идею автоматического выделения информативных признаков из огромного массива изображений, для чего используется высокопроизводительные системы на основе NVidia CUDA [1]. Такие эмпирически выделенные признаки часто показывают существенно лучшие результаты, чем вручную составленные аналоги типа SIFT/SURF/HOG и др. [2].

В недавно вышедшей новой версии фотоорганайзера ZZ Photo для распознавания изображений используется AlexNet-подобная [3] глубокая сверточная сеть, которая заменила иплементированный в прошлые версии приложения детектор Виолы-Джонса для задачи детекции котов и собак [4] и признаки LBP/LPQ [5] для распознавания сцен. С новой технологией качество распознавания возросло в 3-5 раз.

Таблица 1. Качество детекции изображений с домашними любимцами при установленням пороге FAR (False Accept Rate) = 0,5%

	Method name	FRR Error
1	Viola-Jones + LBP + SVM	79,73%
2	Standard “out-of-box” ConvNet, for cats and dogs categories	26,11%
3	Fine-tuned ConvNet + SVM with x^2 kernel	4,35%
4	Reduced and fine-tuned ConvNet + SVM with x^2 kernel (our approach)	6,29%

Исходя из специфики десктопного программного продукта, использование нейронной сети размера ~ 230 Мб неудобно из соображений минимальной требуемой оперативной памяти на ПК пользователя и оптимизации интернет-траффика, поэтому перед нами стала задача минимизации размера модели. В модели AlexNet основное количество настраиваемых весов содержится в последних полносвязных слоях, а не в сверточных слоях, находящихся в начале (Рис. 1, слева).

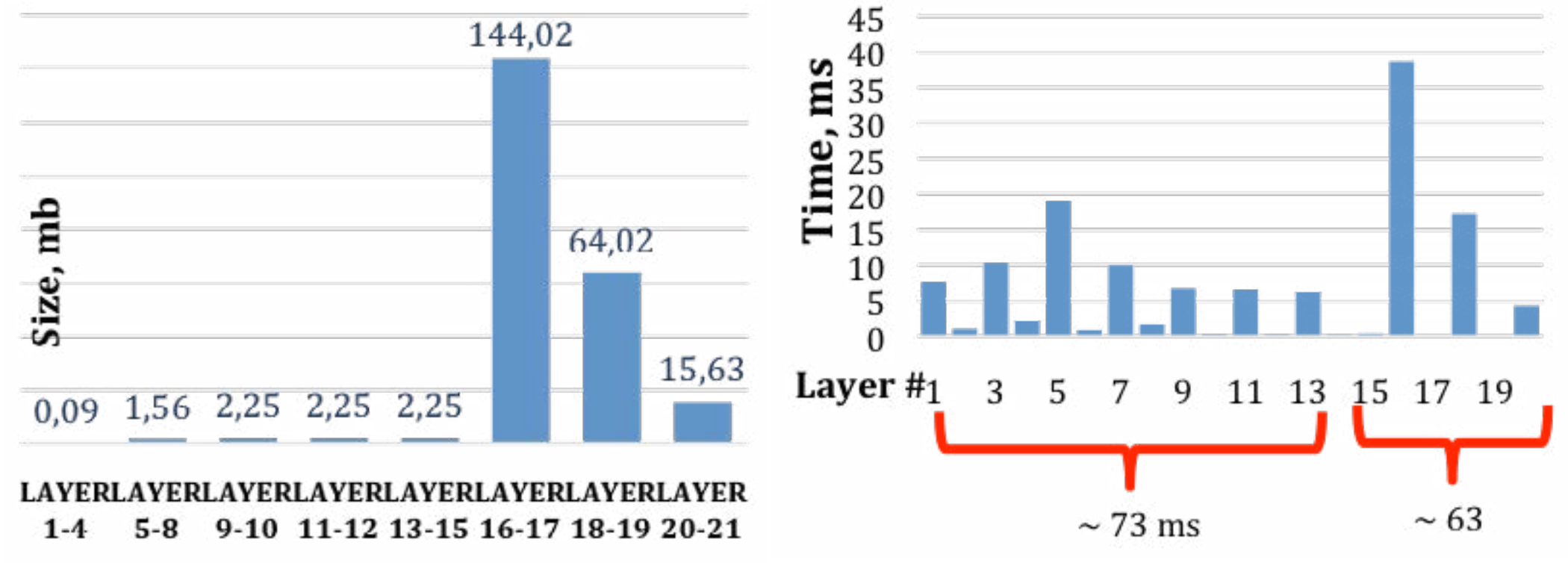


Рис. 1. Слева: Размер весовых коэффициентов в памяти в зависимости от номера слоя. Справа: время работы нейронной сети по слоям для домашнего ПК.

Мы обнаружили, что использование векторов признаков с последнего сверточного (15-го в модели) слоя обеспечивает ошибку детектирования 6,29% , что не намного хуже, чем в полном варианте и вполне приемлемо для нашей задачи. При этом размер нейронной сети драматически уменьшается в 25 раз до 8,97 Мб, а скорость работы – увеличивается примерно в 2 раза (Рис. 1, справа).

Список литературы:

1. Y. LeCun, Y. Bengio, G. Hinton. Deep learning // Nature 521, 436–444 (28 May 2015), doi:10.1038/nature14539.

2. A. S. Razavian, H. Azizpour, J. Sullivan, S. Carlsson. CNN Features off-the-shelf: an Astounding Baseline for Recognition //2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 23-28 June 2014, Columbus, USA, p. 512 – 519.

3. A. Kryzhevsky, I. Sutskever, G.E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks // Advances in Neural Information Processing Systems 25 (NIPS 2012).

4. O. Parkhi, A. Vedaldi, C. V. Jawahar, and A. Zisserman. The Truth about Cats and Dogs // Proceedings of the International Conference on Computer Vision (ICCV), 2011.

5. S. Banerji, A. Verma, C. Liu. Novel Color LBP Descriptors for Scene and Image Texture Classification // Cross Disciplinary Biometric Systems, 2012, 15th International Conference on Image Processing, Computer Vision, and Pattern Recognition, Las Vegas, Nevada, pp. 205-225.

ОЦЕНКА ЭФФЕКТИВНОСТИ РЕАЛИЗАЦИИ АЛГОРИТМА МЕТОДА МОНТЕ-КАРЛО НА СОВРЕМЕННЫХ ГРАФИЧЕСКИХ УСКОРИТЕЛЯХ

А.Н. Ивутин, И.А. Страхов

Подробнее

Увеличение производительности вычислительной техники в последние десятилетия осуществлялось в основном за счет повышения тактовой частоты центрального процессора. Данный способ всегда был и остается наиболее надежным из всех возможных. Однако ввиду технических ограничений при изготовлении интегральных схем уже невозможно рассчитывать на рост частоты работы процессора.

General-purpose graphics processing units (GPGPU) – техника использования графического процессора видеокарты для вычислений общего назначения, которые обычно проводит центральный процессор. CUDA – программно-аппаратная архитектура, разработанная компанией Nvidia для видеокарт серии GeForce 8000 и старше. Платформа обеспечивает набор расширений для языков C/C++/Fortran, позволяющих выражать как параллелизм данных, так и параллелизм задач на уровне мелких и крупных структурных единиц. CUDA использует большое число отдельных нитей для вычислений. Часто каждому вычисляемому элементу соответствует одна нить. Все они группируются в иерархию — grid/block/thread.

Метод «Монте-Карло» – общее название группы численных методов, основанных на получении большого числа реализаций стохастического (случайного) процесса, который формируется таким образом, чтобы его вероятностные характеристики совпадали с аналогичными величинами решаемой задачи.

Актуальность работы заключается в определении оптимальной конфигурации вычислительного ядра для заданного кода, что в дальнейшем позволит создать динамическую модель изменения параметров графического процессора для более эффективного использования ресурсов системы.

В ходе работы была произведена оценка времени нахождения площади треугольника методом Монте-Карло на GPU при различном сочетании блоков в сетке и нитей в блоке.

Для малого количества точек время вычисления велико для каждой из конфигураций. Это связано с невозможностью полностью заполнить все ресурсы графического процессора, что, в свою очередь, приводит к «холостой» работе отдельных скалярных процессоров. При увеличении количества точек «включаются» потоковые возможности мультипроцессоров, что приводит к росту скорости вычисления. Однако для конфигураций с малым числом нитей в блоке с увеличением числа точек время вычисления начинает линейно увеличиваться. Это связано с необходимостью скалярного процессора переключаться множество раз между нитями блока.

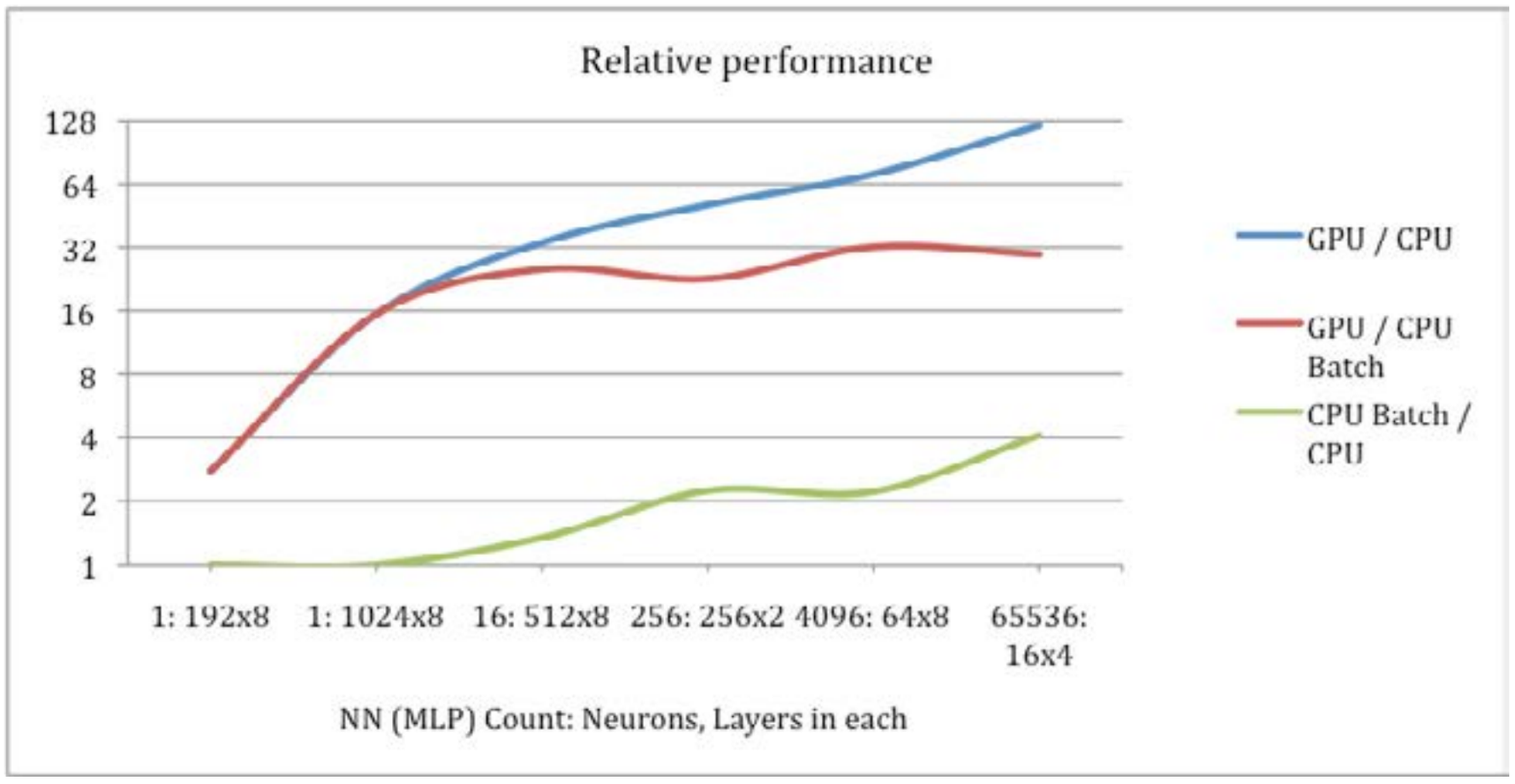
МОДЕЛИРОВАНИЕ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ С ПОМОЩЬЮ ГРАФИЧЕСКОГО АДАПТЕРА ОБЩЕГО НАЗНАЧЕНИЯ

А. А. Королев, А. В. Кучуганов, ИжГТУ им. М. Т. Калашникова

e-mail: ak-hpc@yandex.ru, Aleks_KAV@udm.ru

В рамках работы предложен метод для группировки связей нейронной сети на этапе её проектирования, позволяющий во время обработки сигналов без дополнительных затрат определить, какие связи в нейронной сети могут обрабатываться параллельно. Метод подходит для большого разнообразия архитектур нейронных сетей, в числе которых классический многослойный перцептрон (MLP), свёрточная нейронная сеть (CNN), ограниченная машина Больцмана (RBM), нейронные сети с рекуррентными связями (RNN) и другие. С помощью предложенного метода достигается максимальная степень параллелизма вычислений для рассматриваемой нейронной сети, что позволяет эффективно перенести её обработку на GPU.

В рамках работы был также предложен метод пакетной обработки набора из нескольких нейронных сетей, хорошо согласующийся с упомянутым выше методом группировки. К примеру, при обработке изображений, как правило, приходится «пропускать» через одну и ту же нейронную сеть каждый пиксель или некоторую область изображения. Предложенный метод пакетной обработки позволяет значительно ускорить выполнение за счёт: улучшения локальности данных при обработке на CPU; повышения степени параллелизма при обработке на GPU.



Была реализована библиотека, позволяющая моделировать нейронные сети различных архитектур в режимах на CPU (MSVC++) и на GPU (CUDA) с использованием предложенных методов обработки. Пакетная обработка (Batch) на CPU позволила ускорить обработку в 4 раза, обработка на GPU – в 121 раз. Коэффициент ускорения GPU относительно CPU Batch равен ~32, что примерно соответствует отношению пиковой теоретической производительности GPU GTX 770 к таковой характеристике CPU i5-4670K – использованных при тестировании. Единичная трассировка сигнала по нейронной сети, содержащей 2 105 344 нейронов и 117 964 800 связей, занимает 13.5 мс на GPU GTX 770.

Применение разработанной библиотеки при использовании возможности распознавания образов с помощью свёрточных нейронных сетей на основе автоэнкодеров (кластеризоторов в виде MLP с «узким горлом») позволило сократить время обучения с ~5.75 часов до 31-й минуты с помощью режима GPU (обучающая выборка содержала 60 тыс. образцов). При этом в режиме на CPU без использования пакетной обработки теоретическое время обучения составило бы ~17 часов.

ПОЛЕЗНЫЕ РЕСУРСЫ ПО CUDA

Новый каталог с приложениями, ускоряемыми на GPU можно скачать по ссылке.

Материалы GPU Technology Conference 2015 доступны по ссылке

Форум Разработчиков NVIDIA

Присоединятесь к Форуму CUDA-разработчиков, делитесь своим опытом и узнавайте много нового. http://devtalk.nvidia.com/

Документация по CUDA

Со списком документации по CUDA можно ознакомиться здесь.

Обучение онлайн

Udacity | Coursera | Курс на русском языке

Библиотеки с поддержкой GPU ускорения

Список библиотек с поддержкой GPU ускорения от NVIDIA и партнеров.

GPU Тест-Драйв

Хотите бесплатно протестировать Tesla? Зарегистрируйтесь здесь.

Ускоряйте научные приложения с OpenACC

Протестируйте компилятор PGI OpenACC бесплатно в течение месяца. Подробнее.

Приложения, ускоряемые на GPU

Ознакомиться со списком из более 270 приложений на сайте.

Книги, посвященные CUDA и вычислениям на GPU

Со списком книг, посвященных CUDA и вычислениям на GPU, можно ознакомиться здесь.

Скачайте

CUDA http://developer.nvidia.com/cuda-downloads

Nsight http://www.nvidia.com/object/nsight.html

Страница NVIDIA в vk.com

https://vk.com/nvidia

ВАКАНСИИ CUDA

РАЗРАБОТЧИК ИНСТРУМЕНТАРИЯ ДЛЯ CUDA

NVIDIA

Москва

Комания НВИДИА ищет талантливого инженера для работы над вспомогательными программами для CUDA. Команда занимается разработкой профилировщиков, отладчиков и других программ в интегрированной среде разработки. Вам предстоит заниматься написанием и поддержкой инструментария на C++, QT и Java, а также создавать новые плагины и интегрировать поддержку новых языков и компиляторов для среды Eclipse.

Требования:

- Владение C++, QT
- Опыт программирования на Java как плюс
- Опыт работы с Eclipse IDE как плюсе
- Опыт разработки под Android как плюс
- Разговорный английский язык

Принимаются к рассмотрению только резюме на английском языке.

Подробнее

КОНТАКТЫ

Если вы хотите, чтобы ваша статья появилась в следующем выпуске CUDA Альманах пишите нам на:	Лидия Андреева landreeva@nvidia.com
По вопросам приобретения NVIDIA GPU и по прочим техническим вопросам пишите нам на:	Антон Джораев adzhoraev@nvidia.com