Ученые-биоакустики из Cornell University разработали высокопроизводительную вычислительную платформу для анализа больших данных

Прибор акустического анализа используется в рамках Программы биоакустических исследований для сбора данных о больших усатых китах и других морских млекопитающих. Фото предоставлено Димитрием Пониракисом.

Задача

Детектирование и классификация звуков животных на гигантских объемах акустических данных, собранных в океанах, полях, лесах и джунглях.

Решение

Разработать высокопроизводительную вычислительную платформу для анализа акустических данных на основе MATLAB, Parallel Computing Toolbox и MATLAB Distributed Computing Server.

Результаты

- Сэкономлены годы на разработку
- Время анализа сократилось с недель до часов
- Ранее необработанные данные были проанализированы в течение нескольких дней

Более 30 лет ученые изучают отдельные популяции животных с помощью записи издаваемых ими звуков в океанах, джунглях, лесах и других естественных средах обитания. Эти результаты используются для оценки влияния индустриальных шумов на окружающую среду, отслеживания популяций животных и исследования общения животных друг с другом. Пассивные системы акустического мониторинга записывают звуки непрерывно, что даёт на выходе терабайты данных. Ученые часто неспособны обработать даже 1% этих данных.

Участники Программы биоакустических исследований (BRP) из Лаборатории орнитологии Корнельского университета анализируют огромное количество акустических данных с помощью MATLAB®, Parallel Computing Toolbox™ и MATLAB Distributed Computing Server™. Проект финансируется за счет гранта Управления военно-морских исследований и Национальной океанографической партнерской программы. Во главе его стоят два ведущих исследователя из Корнелла — доктор Кристофер Кларк, главный исследователь и глава BRP, и доктор Питер Дуган, ведущий специалист по обработке данных BRP.

«MATLAB и его инструменты по параллельным вычислениям обладают гибкостью, что дает нам возможность оперативно улучшать и адаптировать наши алгоритмы обработки акустических данных большого размера, — говорит доктор Кларк. — При использовании C++ или аналогичных языков мы не смогли бы передвигаться настолько быстро или исследовать столь много различных сценариев».

Задача

Исследователи, имеющие дело с акустическими данными, должны бороться с шумами, вызванными погодой, другими животными, близкорасположенным оборудованием и транспортом. Изменчивость звуков животных от одной особи к другой также осложняет анализ. Эти два фактора — шум и изменчивость — увеличивают количество ложноположительных и ложноотрицательных результатов, что снижает точность алгоритмов обнаружения.

Обработка сотен терабайт данных, собираемых BRP, представляет собой дополнительную сложность. Типичный проект подразумевает обработку нескольких лет сырых акустических данных — до 10 терабайт — записанных на нескольких каналах. В каждом канале могут улавливаться сотни или миллионы событий-звуков, которые выделяются при представлении данных в спектральном виде. Алгоритмы, проверенные на небольших примерах высокого качества, часто оказываются менее точными при работе с большими, зашумленными данными.

Наконец, средства анализа BRP должны быть способны работать в разных исследовательских программах, средах и в условиях изменяющихся требований. «Ответы на наши исходные вопросы часто приводят к совершенно новым направлениям исследований, и мы должны быть в состоянии обрабатывать эти внезапные изменения в требованиях», — говорит доктор Кларк.

Решение

Специалисты по обработке данных BRP использовали MATLAB для разработки высокопроизводительной вычислительной платформы (HPC) для автоматиче-

«При использовании унаследованного кода на анализ результатов продувки в одной аэродинамической трубе уходило до 40 минут. С помощью MATLAB и GPU время вычисления стало меньше минуты. Для перевода нашего MATLAB-кода на работу с GPU понадобилось 30 минут, т.к. не требовалось низкоуровневого программирования на CUDA», — Кристофер Бар, NASA

ской обработки акустических данных.

Их проект по обнаружению и классификации начался со сбора аудиозаписей целевых животных, записей фонового шума сред их обитания и MAT-файлов архивных акустических данных. Работая в MATLAB, они разрабатывают новые или улучшают существующие алгоритмы по детектированию аудио последовательностей в архивных данных, похожих на те, что есть в каталоге эталонных записей.

В алгоритмах используются поиск по шаблону, выделение границ, анализ связанных областей, свёртка и другие техники из Image Processing Toolbox™, Signal Processing Toolbox™, а также техники машинного обучения из Fuzzy Logic Toolbox™ и Neural Network Toolbox™.

При оценке точности алгоритмов ученые используют Statistics Toolbox™ для вычисления рабочей характеристики приёмника (ROC) и другие кривые производительности.

После отладки и оптимизации алгоритмов на небольших данных с использованием Parallel Computing Toolbox они запускаются на полных архивных наборах данных с использованием MATLAB Distributed Computing Server на кластере с 64 работниками.

Командой BRP был разработан графический интерфейс на MATLAB, в котором исследователь может задать алгоритм, который он хочет использовать, а также количество процессоров и набор данных, с которым будет вестись работа.

BRP в сотрудничестве с Marinexplore и сообществом Kaggle финансировали международный конкурс, на котором более 240 участников представили алгоритмы по обнаружению и классификации нарастающих звуков южных китов. BRP использовали MATLAB HPC-платформу для нахождения наиболее точного алгоритма, который в дальнейшем будет использоваться для предотвращения столкновений морских судов с китами.

Результаты

Сэкономлены годы на разработку. «Прогноз показал, что если бы мы делали все своими силами, это потребовало бы 3 года, 1 миллион долларов и большую внешнюю помощь для разработки необходимой платформы, подобной HPC, — говорит доктор Дуган. — С помощью Parallel Computing Toolbox и MATLAB Distributed Computing Server мы сделали это меньше чем за 3 месяца».

Время анализа уменьшилось с недель до часов. «Для обработки 90 дней записанных данных нашим алгоритмам требовалось 19 недель, — комментирует доктор Дуган. — При использовании Parallel Computing Toolbox и MATLAB Distributed Computing Server тот же анализ на нашем кластере занимает 8 часов».

Ранее необработанные данные были проанализированы в течение нескольких дней. «Один набор содержит 100 000 часов записи звука. Это столь много, что раньше мы обработали менее 1% от них. Оценка показывала, что обработка остальных заняла бы год и более, — говорит доктор Дуган. — На платформе MATLAB HPC мы обработали эти данные 6 раз, с использованием различных алгоритмов детектирования, в течении двух дней».

Области применения

- Цифровая обработка сигналов
- Обработка изображений и видео
- Анализ данных
- Математическое моделирование
- Разработка алгоритмов
- Параллельные вычисления

Промышленность

- Землеведение и океанология

Используемые продукты

- MATLAB
- Fuzzy Logic Toolbox
- Image Processing Toolbox
- MATLAB Distributed Computing Server
- Neural Network Toolbox
- Parallel Computing Toolbox
- Signal Processing Toolbox
- Statistics Toolbox

Дополнительная информация и контакты

Информация о продуктах
matlab.ru/products

Пробная версия
matlab.ru/trial

Запрос цены
matlab.ru/price

Техническая поддержка
matlab.ru/support

Тренинги
matlab.ru/training

Контакты
matlab.ru
E-mail: **matlab@sl-matlab.ru**
Тел.: +7 (495) 232-00-23, доб. 0609
Адрес: 115114 Москва,
Дербеневская наб., д. 7, стр. 8

MathWorks®
Accelerating the pace of engineering and science

mathworks.com