УДК 004.896

УВЕЛИЧЕНИЕ ПРОИЗВОДИТЕЛЬНОСТИ АЛГОРИТМОВ УМЕНЬШЕНИЯ ПРОСТРАНСТВА ДАННЫХ С ПОМОЩЬЮ CUDA И СУПЕРКОМПЬЮТЕРОВ

Костюченко Т.Р.[1], Гончаров Р.Р.[1]
Научный руководитель - кандидат технических наук, доцент Малеев О.Г.[1]
[1]Санкт-Петербургский Политехнический Университет Петра Великого
kostyuchenko.tr@edu.spbstu.ru

Современные системы анализа данных все чаще оперируют многомерными массивами информации, объем которых непрерывно растет. Методы снижения размерности, такие как PCA и t-SNE, являются ключевыми инструментами визуализации и предварительного анализа. Однако их вычислительная сложность – квадратичная и выше – делает прямые расчеты на CPU непрактичными при работе с датасетами, содержащими миллионы объектов. Существующие CPU-реализации, даже с оптимизированными библиотеками линейной алгебры, не обеспечивают приемлемого времени отклика для интерактивных систем. Кроме того, t-SNE обладает принципиальным недостатком: он не позволяет проецировать новые данные без повторного запуска алгоритма на всем массиве, что делает его непригодным для потоковых сценариев.

В своей работе мы исследуем два взаимодополняющих подхода к преодолению указанных ограничений. Был проведен сравнительный анализ эффективности GPU для ускорения вычислений PCA и t-SNE. А также предложено решение проблемы проецирования новых данных без повторного запуска t-SNE. Предлагается гибридный метод, в котором сверточная нейронная сеть обучается аппроксимировать отображение, задаваемое t-SNE, на репрезентативной подвыборке. Такой подход сочетает высокое качество визуализации исходного алгоритма с оперативностью.

Эксперименты проводились на двух типах оборудования. Для сравнения CPU и GPU использовалась локальная станция. Для нейросетевой аппроксимации применялись узлы tornado-k40 суперкомпьютерного кластера «Политехник – РСК Торнадо.

Программный стек включал CPU-реализации из библиотеки Scikit-learn 1.3, GPU-реализации из cuML 23.08, обработку данных из библиотек Pandas и NumPy, визуализацию из библиотеки Matplotlib и управление памятью из библиотеки CuPy.

В работе использовались два набора данных. Тестовый датасет DIGITS служил для визуальной проверки качества алгоритмов. Основной датасет содержит 121 параметрический признак, общий объем около 4ГБ и применялся для измерения производительности.

В ходе эксперимента было подсчитано, что перенос основного датасета в оперативную память занял 0.43 с., тогда как загрузка в видеопамять GPU потребовала всего 0.043 с., что даёт ускорение в 10 раз. Далее для PCA и t-SNE фиксировались параметры: для t-SNE значение perplexity = 30, число итераций = 1000, random_state = 42; для PCA число компонент = 2. Объём выборки варьировался от 6000 до 60000 записей. Измерялось полное время выполнения, включая передачу данных на GPU.

GPU реализация для PCA начинает превосходить CPU только при объёме выборки более 18000 записей. При 60000 образцов ускорение составляет около 3 раз. В противоположность этому, t-SNE на GPU демонстрирует значительный выигрыш уже на малых объёмах: при 6000 записях время сокращается с 6.84 с. на CPU до 1.25 с. на GPU. При увеличении выборки до 60000 время на GPU составляет 3.09 с. против 77.32 с. на CPU, что соответствует ускорению в 25 раз.

Результаты снижения размерности на тестовом датасете DIGITS подтверждают известные свойства алгоритмов. PCA сохраняет глобальную структуру, но кластеры цифр

частично перекрываются. t-SNE формирует чётко разделённые компактные группы, демонстрируя высокое качество сохранения локальных отношений.

Переходя к нейросетевой аппроксимации, получаем, что нейросеть точно воспроизводит взаимное расположение кластеров: формы и относительные позиции групп сохраняются, границы лишь незначительно «размываются». Это свидетельствует о том, что сеть успешно выучила нелинейное отображение, задаваемое t-SNE.

При переходе к основному высокоразмерному датасету двумерная проекция, полученная прямым t-SNE, является сложной фрагментированной структурой.

В ходе исследований получены временные характеристики для трёх фрагментов датасета разного объёма. Время предсказания нейросети на 2–3 порядка меньше прямого расчёта t-SNE. Даже с учётом затрат на обучение, при многократном использовании модели общий выигрыш достигает сотен и тысяч раз.

Полученные данные подтверждают неоднородную эффективность GPU для разных алгоритмов. t-SNE с его квадратичной сложностью и массовым параллелизмом (независимые попарные расстояния) является идеальным кандидатом для GPU ускорения. Напротив, PCA содержит последовательные этапы, что ограничивает выигрыш от параллелизации; GPU становится эффективным лишь при достаточно больших размерах задачи. На основе анализа определены критические точки. Для t-SNE использование GPU рекомендуется всегда, начиная с малых объёмов. Для PCA выбор платформы должен определяться размером выборки: при менее 5000 записей целесообразен CPU, при более 5000 GPU.

Разработанный метод нейросетевой аппроксимации позволяет преодолеть главный недостаток t-SNE - невозможность проецирования новых точек без пересчёта. Единожды обученная на репрезентативной подвыборке сеть способна за микросекунды выдавать координаты для произвольного числа новых объектов, сохраняя кластерную структуру. Ограничения связаны с необходимостью подбора архитектуры и гиперпараметров, а также с потенциальным падением точности при сильном несоответствии распределения новых данных обучающей выборке.

Несмотря на существование разнообразных подходов к оптимизации вычислений, включая перспективные методы аппаратно-ориентированного дизайна, такие как бинарные нейронные сети [1], при анализе больших данных наиболее актуальным остается использование GPU и экосистемы CUDA. Сравнение с современными исследованиями подтверждает этот тезис: полученные ускорения t-SNE на GPU согласуются с результатами, опубликованными в документации NVIDIA cuML, где заявлено ускорение в 10-50 раз на GPU [2]. Хотя идея аппроксимации t-SNE нейросетью также встречается в литературе (parametric t-SNE [3]), большинство работ ограничиваются малыми датасетами. Наше исследование демонстрирует работоспособность подхода на данных объёмом в несколько гигабайт и оценивает реальный выигрыш во времени, включая этап обучения.

Практическая значимость работы заключается в выработке эмпирических правил, для оптимизации ресурсов при анализе данных: для t-SNE на средних и больших датасетах рекомендуется GPU; для PCA зависит данных; при частом поступлении новых данных эффективно применять гибридный подход.

Литература

1. Королев, Д. О. Исследование эффективности применения бинарных нейронных сетей при детектировании объекта на изображении / Д. О. Королев, О. Г. Малеев // Информатика и ее применения. – 2023. – Т. 17, № 3. – С. 88-92. – DOI 10.14357/19922264230312. – EDN TOCVAL.
2. NVIDIA cuML Documentation. [Электронный ресурс]. Режим доступа: https://docs.rapids.ai/api/cuml/stable/ (дата обращения: 10.12.2025).
3. Van Der Maaten L. Accelerating t-SNE using tree-based algorithms // Journal of machine learning research. – 2014. P. 14-20.