УДК 004.912 004.021 004.822

Воронков И.М.

Московский физико-технический институт (государственный университет)

Использование графических ускорителей для задач текстовой обработки с применением нейросетевых технологий

Последние несколько лет активно развиваются алгоритм обработки текстовой информации на основе нейронных сетей. В докладе рассматривается обобщение алгоритма на основе нейронной сети Хопфилда, которая является частным случаем сети прямого распространения, на случай не бинарного выхода нейрона [1] и реализация вычислений на графических платах.

Обработка текстовой информации с использованием нейросетевой технологии TextAnalyst [2], реализованная на персональном компьютере, имеет ограничения по объему обрабатываемой информации, как вследствие ограничений объема оперативной памяти, так и вследствие ограничений по быстродействию. Эти ограничения возникают в процедуре итеративной перенормировки весовых коэффициентов понятий ассоциативной сети. Применение аппаратных ускорителей позволяет снять эти ограничения, в некоторых случаях значительно поднять потолок по объему обрабатываемой информации.

Обозначив вершинами сети понятия, извлеченные из текста на этапе частотной обработки, получаем, что степень связанности *i*-го понятия с *j*-м проецируется в вес связи от *i*-го нейрона. В то же время частота встречаемости понятия в тексте будет соответствовать изначальному уровню возбуждения соответствующего нейрона. Поэтому, при анализе используется следующий алгоритм:

1. Производится инициализация значений весовых характеристик вершин сети.
2. Итерационно обновляются веса нейронов:
   a. Для каждого нейрона вычисляется входная сумма $S_i = \sum_{j=1}^{N} w_{i,j} \cdot x_j^t$.
   b. Находится среднее значение суммы $M = \frac{1}{N} \sum_i S_i$.
   c. Вычисляется новый вес нейрона $x_i^{t+1} = \sigma\left(\frac{S_i}{M}\right)$, где $\sigma(x) = \frac{A}{1 + e^{B \cdot (1-x)}}$, $A = 100$, $B \in (0...5.0)$ - являются параметрами алгоритма.
   d. Вычисляется новое значение энергии сети *E*.

e. Процесс завершается после выполнения заданного числа циклов, либо при выполнении условия сходимости: $\left|E^{t+1} - E^{t}\right| < \delta$

3. Результатом обработки является значение весов нейронов, характеризующих относительную значимость понятий в тексте.

Для использования графического процессора в качестве ускорителя вычислений в основной программе вместо исходного кода, реализующего вышеописанный алгоритм, вставляется программный код обращения к специальному потоку вычислений, в котором организован цикл управления вычислениями на графическом процессоре. Так как графические процессоры различных производителей (NVIDIA и ATI) предоставляют возможность программирования вычислений общего назначения GPGPU с помощью собственных программных средств, то в результате подобная организация вычислений позволяет в одном приложении организовать поддержку аппаратно зависимых вычислений путем подключения динамических библиотек, в которых реализованы потоки управления графическим процессором. Эти динамические библиотеки компилируются с использованием средств программирования производителя графических процессоров.

Так как в случае обработки нескольких текстовых документов процесс перенормировки нейронной сети происходит несколько раз и происходит рост сети, то для оптимизации вычислений на графической плате выделяется область памяти, в которую копируется входные данные нейронной сети, а также область памяти для результата и некоторых промежуточных результатов. Выделение памяти организовано некоторыми фиксированными по размеру частями, для удовлетворения требований по выравниванию данных, что позволяет производить обработку максимально параллельно и без возникновения конфликтов. Из оперативной памяти происходит копирование только новых входных данных, а результаты предыдущей обработки являются также и частью входных данных. Таким образом, уменьшается время на обмен между оперативной памятью и памятью графического процессора, что позволяет получить более высокие показатели быстродействия при пакетной обработке массива документов, а также происходит ускорение вычислений.

СПИСОК ЛИТЕРАТУРЫ

1. *Харламов А.А*. Нейросетевая технология представления и обработки информации (естественное представление знаний). - М.: «Радиотехника», 2006.

2. *Харламов А.А., Ермаков А.Е., Кузнецов Д.М.* TextAnalyst - комплексный нейросетевой анализатор текстовой информации. Вестник МГТУ им. Н.Э. Баумана. N 1, 1998. - С. 32-36