ОПТИМИЗАЦИЯ ПРИМЕНЕНИЯ ПРИЛОЖЕНИЙ TENSORFLOW НА РАБОЧЕЙ СТАНЦИИ INTEL® XEON® PLATINUM

С.К. Шикота

Научный Центр РАН в Черноголовке, 142432 Черноголовка, Россия

E-mail: sveta@chg.ru

Платформа TensorFlow является одним из наиболее развитых наборов программных продуктов с открытым кодом для задач машинного обучения. С другой стороны, рабочие станции на базе процессоров Intel® Xeon® Scalable® представляются перспективным аппаратным решением для задач машинного обучения. Их отличительная черта состоит в комбинации из трех важных элементов. Во-первых, это большое число тяжелых ядер в одном CPU, более двух десятков. Во-вторых, это наличие двух устройств AVX-512 (Advanced Vector Extension 512), которые дают возможность работы с 512-битными регистрами. В третьих, это очень большой размер памяти, на одной материнской плате 1.5 TB высокоскоростной памяти DDR4, которая поддерживается большим кэшем второго уровня. Такое устройство одновременно предоставляет большую скорость вычислений и работу с данными большого объема в оперативной памяти. В перспективе, это позволяет проводить анализ сложных проблем с большим объемом данных. Мы приводим результаты проверки производительности некоторых пакетов распознаниz изображений из платформы TensorFlow.

Ключевые слова: оптимизация приложений, архитектура Intel® Xeon® Scalable, ускорители вычислений, машинное обучение

1. Введение

Большой интерес исследователей в настоящее время привлекает возможность использования методов машинного обучения для получения нового знания в различных областях естественных и социальных наук [1]. Платформа TensorFlow [2] является одним из наиболее развитых наборов программных продуктов с открытым кодом для решения задач методами машинного обучения. Решение таких задач требует серьезных аппаратных затрат как по объему анализируемой информации, так и по объему проводимых вычислений. Для этой цели платформа TensorFlow использует аппаратные решения на базе ускорителей вычислений Nvidia [3]. В конце 2017 года на рынке появились серверные решения на базе новой архитектуры Intel® Xeon® Scalable® [4]. Применение такой архитектуры позволяет надеяться на существенное ускорение работы сервера в целом. Для такой надежды есть три причины, которые основаны на важных усовершенствованиях архитектуры Intel® Xeon® Scalable®. Во-первых, это большое число тяжелых ядер в одном CPU, более двух десятков. Это позволяет обрабатывать больший объем информации в единицу времени за счет большого числа параллельных потоков. Во-вторых, наличие в каждом ядре двух устройств AVX-512 (Advanced Vector Extension 512), которые дают возможность работы с 512-битными регистрами. Это позволяет одновременно производить в архитектуре SIMD (Single Instruction Multiple Data) обработку восьми 64-битных чисел с плавающей запятой на каждом из двух устройств FMA [5]. Теоретически это позволяет ускорить вычисления над 32-битными целыми числами в 16 раз. Такое увеличение поддерживается также наличием у каждого процессора ускоренного обмена с оперативной памятью за счет работы шести каналов обмена с памятью. В третьих, скорость вычислений поддерживается быстрой (2666 MHz DDR4) оперативной памятью в 758 GB. Общая память для двух процессоров на одной материнской плате узла может достигать 1.5 TB. Поддержка этих функция реализована в библиотеке Intel® Math Kernel Libraries (Intel® MKL) [6].

Такие характеристики рабочей станции позволяют надеяться на существенное ускорение вычислений с данными большого объема.

2. Постановка эксперимента

Для проведения эксперимента мы использовали сервер с максимально возможной оперативной памятью 2x768 GB, 2666MHz DDR4 и с двумя процессорами Intel(R) Xeon(R) Platinum 8164 (GenuineIntel processor 103, cpu family 6, model 85, 2.00 GHz). Каждый процессор имеет 26 доступных для вычисления ядер, что с учетом hyperthreading позволяет обрабатывать информацию и одновременно запускать 104 нити на одном узле.

Были установлены следующие версии системного и программного обеспечения:

- Операционная система CentOS Linux release 7.5.1804;
- Компилятор gcc release 4.8.5;
- Компилятор Intel C++ и библиотеки release 2018.3;
- OpenMP 4.5 (with the SIMD Directives);
- TensorFlow release 1.10.0;

Тестировались следующие сети:

- VGG16 – 16-ти слойная нейронная конволюционная сеть, разработана коллективом Visual Geometry Group (университет Оксфорда) для распознавания элементов на сложных изображениях [7];
- Inception v3 – 16-ти - 19-ти слойная конволюционная нейронная сеть, разработана Google [8];
- Resnet50 – 50-ти слойная остаточная нейронная сеть, разработана Microsoft [9].

Такой выбор приложений позволяет нам провести сравнение с их исследованием на другой реализации архитектуры Intel® Xeon® Scalable, с 28-ми ядерном процессоре серии Platinum 8180 [10] со следующими параметрами 2x188 GB, 2666MHz DDR4 и с двумя процессорами Intel(R) Xeon(R) Platinum 8180 (cpu family 6, model 85, 2.50 GHz). Операционная система CentOS Linux release 7.4.

При использовании AVX-512 и, в особенности, набора команд FMA, оба процессора понижают частоту– интенсивное выполнение команд AVX-512 ведет к выделению большого количества тепла, и понижение частоты препятствует перегреву процессора. К сожалению, информация о реальной частоте при проведении эксперимента нам оказалась по ряду причин недоступной. В работе [10] такая информация также не приведена.

3. Результаты эксперимента

Для проведения эксперимента мы выбрали оптимальные для нашего оборудования параметры. Число нитей было выбрано равное числу физических ядер на каждом процессоре, всего нитей 52. Для сети vgg16 мы выбрали одну связывающую нить, а для двух других сетей – две. Это соответствует максимальным установкам в эксперименте с процессором 8180 с большим числом нитей – по 28 на процессор, всего 56 нитей на узел. В последней колонке таблицы приведены результаты нашего эксперимента. Отличие результатов от работы [10] можно объяснить двумя обстоятельствами – 1) меньшим числом нитей, 2) большей частотой процессора 8180 по сравнению с процессором 8164. В то же время, данные отличаются сильнее, чем можно ожидать от этих двух обстоятельств, что можно объяснить тем, что авторы [10] провели оптимизацию некоторых процедур. Нам такая оптимизация пока недоступна – нет достаточно подробных описаний деталей функционирования AVX-512/FMA.

	Размер пакета	изображений в сек [10]	изображений в сек
Vgg16	128	44	32
Inception v3	64	58	52
Resnet50	128	90	70

4. Заключение

Исследования показывают, что линейка процессоров Intel® Xeon® Platinum может повысить эффективность использования методов машинного обучения на задачах распознавания изображений.

Замедление обработки изображений при использовании одного процессора и одной нити составляет примерно 50, что указывает на высокую эффективность использования новой архитектуры для задач глубокого машинного обучения.

Мы планируем дальнейшие исследования для более точного установления коэффициента масштабирования и анализа аппаратных особенностей (поиск влияния границ памяти, изучение влияния числа нитей, поиск узких мест и т.п. за счет варьирования размера изображений, числа нитей и т.д.).

Благодарности

Работа выполнена в рамках гранта РФФИ 17-07-01377.

Список литературы

[1] Nicke M. et al. A Review of Relational Machine Learning for Knowledge Graphs // Proceedings of the IEEE 2015, vol. 104, No. 1, pp. 11 – 33, DOI: 10.1109/JPROC.2015.2483592

[2] Сайт проекта TensorFlow. [электронный ресурс]. Дата обновления: 09.10.2018. URL: https://www.tensorflow.org/ (дата обращения 04.112018)

[3] Сайт графического ускорителя Nvidia V100. [электронный ресурс]. URL: https://www.nvidia.com/ru-ru/data-center/tesla-v100/ (дата обращения 04.11.2018)

[4] Сайт процессоров Intel® Xeon® Scalable®. [электронный ресурс]. URL: https://www.intel.com/content/www/us/en/processors/xeon/scalable/xeon-scalable-platform.html (дата обращения 04.11.2018)

[5] AVX-512. [электронный ресурс]. URL: https://ru.wikipedia.org/wiki/AVX дата обращения 04.11.2018)

[6] Сайт библиотеки программ Intel® Math Kernel Libraries (Intel® MKL). [электронный ресурс]. URL: https://software.intel.com/en-us/mkl (дата обращения 04.11.2018)

[7] Simonyan K., Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition // arXiv:1409.1556

[8] Szegedy C. et al. Rethinking the Inception Architecture for Computer Vision // arXiv:1512.00567

[9] Kaiming He et al. Deep Residual Learning for Image Recognition // arXiv:1512.03385

[10] Vivek Rane, AG Ramesh. TensorFlow* Optimizations for the Intel® Xeon® Scalable Processor. [электронный ресурс]. URL: https://ai.intel.com/tensorflow-optimizations-intel-xeon-scalable-processor/ (дата обращения 06.11.2018)

OPTIMISATION OF TENSORFLOW APPLICATIONS ON THE WORKSTATION INTEL® XEON® PLATINUM

S. Shikota

Science Center in Chernogolovka, 142432 Chernogolovka, Russia

E-mail: sveta@chg.ru

TensorFlow is one of the most developed software with the open code является in the framework of the machine learning. At the same time, the workstations with the Intel® Xeon® Scalable® architecture looks the quite promising tools for the machine learning. The highlights of the architecture are in the three important features. Firstly, it is a big number of heavy cores in the CPU, typically more than twenty. In our case, it is 26 cores in each of two CPU 8164, and hyperthreading in addition. Secondly, it is two AVX-512 (Advanced Vector Extension 512) per core, which provides 512-bit registers. Theoretically, it gives acceleration with coefficient 16 while operating with 32-bit numbers. Thirdly, it is a huge RAM, with 1.5 TB on the board. Altogether, the features lead to fast calculations and good performance with the Big Data manipulations. We discuss the performance results of some applications from the TensorFlow.

Keywords: application optimization, architecture Intel® Xeon® Scalable, computing accelerators, machine learning