УДК 004.8

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ПРОИЗВОДИТЕЛЬНОСТИ И ЭКОНОМИЧЕСКОЙ ЭФФЕКТИВНОСТИ ВИДЕОКАРТ ПРИ РАБОТЕ С МНОЖЕСТВОМ МОДЕЛЕЙ BERT

Бадретдинов Даниил Владимирович, Зайченко Елена Аркадьевна
Белорусско-Российский университет, г. Могилев, Беларусь
daniilbadret@gmail.com, helena.zai@mail.ru

Аннотация. В работе проведён сравнительный анализ производительности и экономической эффективности графических ускорителей (GPU) различных архитектур и ценовых категорий при работе с большим количеством моделей BERT. Исследованы GPU серии RTX, A100, H100 и других, выявлены модели с наилучшим соотношением производительности к стоимости. Установлено, что наибольшая память и высокая стоимость GPU не всегда обеспечивают оптимальную экономическую эффективность. Полученные результаты позволят сделать обоснованный выбор видеокарт для задач обработки естественного языка.

Ключевые слова: BERT, трансформеры, GPU, видеокарты, производительность, экономическая эффективность, искусственный интеллект, глубокое обучение, нейронные сети, NLP.

COMPARATIVE ANALYSIS OF THE PERFORMANCE AND COST-EFFECTIVENESS OF VIDEO CARDS WHEN WORKING WITH MULTIPLE BERT MODELS

Badretdinov Daniil Vladimirovich, Zaichenko Elena Arkad'evna
Belarusian-Russian University, Mogilev, Belarus
daniilbadret@gmail.com

Abstract. The paper provides a comparative analysis of the performance and cost-effectiveness of graphics accelerators (GPUs) of various architectures and price categories when working with a large number of BERT models. GPUs of the RTX, A100, H100 and others series have been investigated, and models with the best performance-to-cost ratio have been identified. It has been found that the largest memory and the high cost of a GPU do not always provide optimal economic efficiency. The results obtained will make it possible to make an informed choice of video cards for natural language processing tasks.

Keywords: BERT, transformers, GPU, graphics cards, performance, economic efficiency, artificial intelligence, deep learning, neural networks, NLP.

Модели семейства BERT, построенные на архитектуре трансформера, активно применяются в задачах обработки естественного языка [1], однако их использование требует значительных вычислительных ресурсов. Это обуславливает необходимость поиска оптимального графического оборудования, обеспечивающего высокую производительность при приемлемых финансовых затратах. В работе исследуется вопрос о том, всегда ли высокая стоимость и большие технические характеристики GPU гарантируют соответствующий рост эффективности при работе с множеством моделей BERT.

Цель работы – провести сравнительный анализ производительности различных видеокарт при работе с большим количеством моделей BERT, выявить зависимость эффективности от технических характеристик и стоимости оборудования, а также определить видеокарты, обеспечивающие оптимальное соотношение «производительность/стоимость».

В ходе исследования была произведена серия замеров производительности различных графических ускорителей при выполнении множества экземпляров моделей семейства BERT. Для тестирования использовался пользовательский скрипт на базе библиотеки PyTorch, последовательно выполняющий инференс 1–19 идентичных моделей, загруженных в память одной видеокарты. Для всех моделей использовался одинаковый входной текст, а результат измерялся как суммарное время прохождения всех моделей в одном цикле без обучения.

В ходе замеров фиксировались значения времени выполнения прогонов моделей на следующих GPU: RTX 6000 Ada, RTX 5000 Ada, RTX A6000, NVIDIA A100 SXM4, H100 NVL, A40, L40S, Tesla V100, Quadro RTX 6000, RTX 4090, RTX 4070 Laptop и GTX 1070.

По результатам замеров было установлено, что наименьшее среднее время выполнения одного прогона модели при 1–10 экземплярах показали видеокарты RTX 6000 Ada, L40S, H100 NVL и RTX 4090 (таблица 1). При этом H100 NVL, несмотря на самую высокую цену аренды и максимальную память, не продемонстрировала существенного преимущества по сравнению с более доступными по цене RTX 6000 Ada и L40S.

Видеокарты, такие как Tesla V100, Quadro RTX 6000, GTX 1070, имели существенно более высокое время выполнения, особенно при увеличении количества одновременно исполняемых моделей. Это указывает на их ограниченность для современных задач масштабируемого инференса [2].

Анализ зависимости между техническими характеристиками и производительностью показал:

1) Объём видеопамяти позволяет запускать большее число моделей, но напрямую не определяет скорость их выполнения. Например, A100 SXM4 с 40

ГБ памяти проигрывает RTX 6000 Ada с тем же объёмом по производительности.

2) Вычислительная мощность (TFLOPS) влияет на скорость, но не является абсолютным индикатором: RTX 6000 Ada и L40S имеют одинаковую мощность, однако L40S стоит дешевле и демонстрирует схожие результаты [3].

Таблица 1. Сравнение времени инференса BERT на исследуемых GPU

Кол-во моделей	RTX 6000Ada	RTX 5000Ada	RTX A6000	A100 SXM4	H100 NVL	A40	L40S	Tesla V100	Q RTX 6000	RTX 4090
1	0,0101	0,0165	0,0219	0,0240	0,0093	0,0202	0,0102	0,0309	0,0260	0,0179
2	0,020	0,0335	0,0462	0,0491	0,0187	0,0403	0,0208	0,0588	0,0529	0,0377
3	0,0317	0,0487	0,0709	0,0756	0,0277	0,0605	0,0318	0,092	0,0967	0,0540
4	0,0426	0,0655	0,0962	0,1005	0,0367	0,0812	0,0425	0,1306	0,1081	0,0731
5	0,0546	0,0828	0,1128	0,1279	0,0481	0,1026	0,055	0,1535	0,1308	0,091
6	0,0652	0,0995	0,1404	0,1519	0,0576	0,1235	0,0621	0,2227	0,1962	0,1100
7	0,0763	0,1152	0,1743	0,1774	0,069	0,1439	0,0739	0,2252	0,1892	0,1418
8	0,0880	0,1311	0,1808	0,2060	0,0793	0,1653	0,0865	0,2542	0,2120	
9	0,0985	0,1602	0,2093	0,2276	0,0889	0,1897	0,1008	0,2910		
10	0,1100	0,1662	0,2411	0,2554	0,098	0,2093	0,1120	0,3727		
11	0,1222	0,1804	0,2686	0,2762	0,1085	0,2268	0,1203	0,3380		
12	0,1320	0,1990	0,2781	0,3034	0,1176	0,2489	0,1315	0,3796		
13	0,1441		0,3077	0,3292	0,1262	0,2713	0,1436			
14	0,1496		0,3426	0,3574	0,139	0,2898	0,1526			
15	0,1666		0,3449	0,3814	0,1493	0,312	0,1641			
16	0,1766		0,3822		0,1603	0,335	0,1733			
17	0,1876		0,4338		0,1690	0,3527	0,1850			
18	0,1990		0,4117		0,1791					
19	0,2088		0,4559		0,1927					

Также были собраны характеристики каждой видеокарты: заявленная вычислительная мощность (TFLOPS), стоимость аренды в долларовом эквиваленте за час, объём видеопамяти (таблица 2).

По наблюдениям, наилучшее соотношение производительности и стоимости аренды продемонстрировали видеокарты L40S и RTX 6000 Ada. Они обеспечили высокую скорость выполнения при сравнительно невысокой стоимости аренды, что делает их особенно привлекательными для масштабируемых задач с большим числом моделей.

В то же время, такие решения, как H100 NVL, хоть и показали хорошие результаты по времени выполнения, оказались значительно дороже, не обеспечивая пропорционального прироста эффективности. Аналогично, видеокарты A100 SXM4 и A6000 уступили по общей выгоде более доступным GPU, что подтверждает, что высокая цена и большая видеопамять не всегда означают лучшую эффективность.

Таблица 2. Характеристики исследуемых видеокарт

Характеристика	RTX 6000Ada	RTX 5000Ada	RTX A6000	A100 SXM4	H100 NVL	A40	L40S	Tesla V100	Q RTX 6000	RTX 4090
Мощность, TFLOPS:	81,4	63,6	36,1	15,6	48,3	29,9	73,3	11,3	11,9	81,4
Цена, $/hr:	0,67	0,556	0,514	0,655	3,206	0,422	0,606	0,201	0,236	0,316
Объём видеопамяти, GB:	48	32	48	40	93,6	45	45	32	22,5	24

Для более объективной оценки эффективности видеокарт был введён индекс производительности, рассчитываемый как отношение теоретической вычислительной мощности (в терафлопсах) к произведению стоимости аренды в долларах за час и среднего времени инференса одной модели BERT в секундах.

Этот индекс показывает, насколько выгодно использовать ту или иную видеокарту при выполнении масштабируемого инференса: чем выше индекс, тем лучше соотношение производительности к затратам. Среднее время инференса рассчитывалось как среднее арифметическое времени запуска всех протестированных экземпляров моделей BERT (таблица 3).

Таблица 3. Индексы эффективности видеокарт (BERT-инференс)

Видеокарта	RTX 6000Ada	RTX 5000Ada	RTX A6000	A100 SXM4	H100 NVL	A40	L40S	Tesla V100	Q RTX 6000	RTX 4090
Индекс	1107	1057	295	117	152	379	1234	265	398	3429

По результатам расчётов наибольшие значения индекса показали RTX 4090, RTX 6000 Ada и L40S, что подтверждает их высокую эффективность при сравнительно умеренной стоимости аренды. Напротив, такие видеокарты как

H100 NVL и A100 SXM4, несмотря на высокие технические характеристики, продемонстрировали более низкий индекс из-за значительно большей стоимости аренды.

В ходе проведённого исследования была проанализирована производительность и экономическая эффективность различных видеокарт при выполнении большого количества моделей BERT. Эксперименты показали, что высокие технические характеристики и стоимость оборудования не всегда гарантируют пропорциональное увеличение производительности.

Видеокарты RTX 6000 Ada и L40S продемонстрировали наилучший баланс между скоростью выполнения и стоимостью аренды, что делает их оптимальными для задач, связанных с массовым запуском моделей BERT. В то же время, более дорогие решения, такие как H100 NVL и A100 SXM4, не обеспечили соответствующего прироста эффективности, несмотря на значительно большую цену и объём памяти.

Таким образом, выбор видеокарты для практических задач, связанных с задачами обработки естественного языка с помощью модели семейства BERT должен основываться не только на максимальных технических показателях, но и на соотношении «производительность/стоимость» в конкретных сценариях использования.

Источники

1. Новиков, А. С. Использование языковой модели BERT для анализа текстов на русском языке / А. С. Новиков, Е. В. Шарлаев // Наукосфера. 2021. № 6-1. С. 200-202.

2. Вечканова, Ю. С. Использование BERT-моделей естественных языков для управления нормативно-справочной информацией / Ю. С. Вечканова, С. А. Федосин // Огарёвские чтения: Материалы всероссийской с международным участием научной конференции. Том Часть 1. Саранск: Национальный исследовательский Мордовский государственный университет им. Н.П. Огарёва, 2022. С. 416-419.

3. Прошина, М. В. Оптимизация больших языковых моделей / М. В. Прошина, А. Н. Виноградов // Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем: Материалы Всероссийской конференции с международным участием. Москва: Российский университет дружбы народов им. П. Лумумбы, 2024. С. 231-237.