GPU リソースを制限した場合の LLM 推論の性能評価

高頭　まどか　(指導教員：小口　正人)

1　はじめに

近年，ChatGPT をはじめとする大規模言語モデル（LLM）の驚異的な能力に注目が高まっている．自然言語理解や自然言語生成，知識集約的なタスク，推論能力に優れ，多数の実践的な応用例が示されている [1]．これに伴い，LLM の学習や推論処理を高速化することのできる GPU の需要が急増し，調達や利用コストの増加が問題になっている．そこで，GPU の利用効率を高めて相対的にコストを下げることが重要になってくる．しかし，多くの環境で GPU 使用率は低く，GPU を十分に活用しきれていない現状がある．そこで本研究では，GPU リソースの効率的な利用に向け，モデルごとに最低限必要なリソース量を明らかにし，最適なリソース割り当てを実現することを目指す．今回は，LLM 推論を実行する際に GPU リソースを制限してスループット等の性能測定を行い，各モデルに必要なリソースと性能の関係について調査した．

2　実験

2.1　実験概要

NVIDIA が提供するオープンソース推論フレームワークである Triton Inference Server と LLM 推論をローカルで高速に実行するためのオープンソースライブラリである vLLM を組み合わせて LLM 推論を実行した．その際，vLLM の Engine Argument である gpu-memory-utilization[2] を制御することで，使用可能な GPU メモリ量に制限を設けた．gpu-memory-utilization は 0 から 1.0 までの範囲で指定でき，デフォルト値は 0.9 となっている．gpu-memory-utilization: 0.9 のときは総 GPU メモリ量のうち 90%が使用可能となる．

今回は表 1 に示す 3 つのモデルを使用し GPU メモリ量を制限した状態で LLM 推論を実行した．そして Triton Inference Server 上で動作するモデルの推論性能を評価する CLI ツールである Perf Analyzer[3] で，単位時間あたりの推論実行数である Throughput (infer/sec) の測定を行った．実験環境を表 2 に示す．GPU は NVIDIA Tesla V100 が 2 基搭載されている．

表 1: 実験対象のモデル

モデル名	Developer	パラメータ数
gpt2-small	OpenAI	0.117B
Llama3.2-1B	Meta	1B
Llama3.2-3B	Meta	3B

表 2: 実験環境

実験用サーバ	PRIMERGY CX2570 M1
CPU	Intel Xeon CPU E5-2697v3 2.60GHz (x 2)
CPU メモリ	DDR4-2133 128GB
GPU	NVIDIA Tesla V100-PCIE (x 2)
GPU メモリ	HBM2 16GB (/GPU)
OS	Rocky Linux 9.4

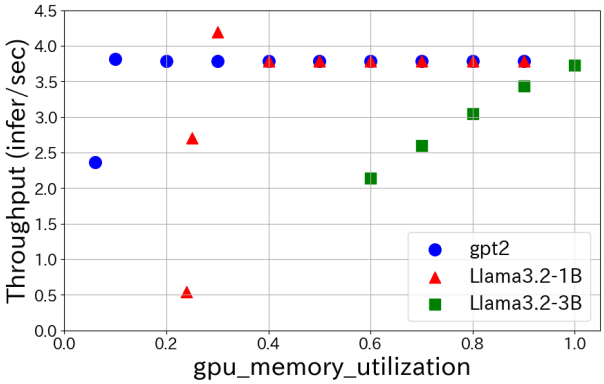

図 1: gpu-memory-utilization を変えたときの Throughput の変化

2.2　実験結果

2.2.1　GPU1 基使用

まずは GPU を 1 基のみ使用して推論を実行した．3 つのモデル (gpt2・Llama3.2-1B・Llama3.2-3B) で gpu-memory-utilization を変えたときの Throughput の変化を図 1 に示す．モデルのパラメータ数が 3 つの中で最小である gpt2 は gpu-memory-utilization: 0.1 以上のとき Throughput はほぼ横ばいであり，GPU メモリを制限することによる性能の低下がほぼ見られないということがわかった．次にパラメータ数の多い Llama3.2-1B は gpu-memory-utilization: 0.24 以上で実行可能となった。そして、gpu-memory-utilization: 0.3 のときに Throughput がピークとなる 4.19 infer/sec を記録し，それ以上 GPU メモリ使用量を増やしても Throughput が向上することはなかった．上記の 2 つのモデルは，gpu-memory-utilization が大きいほど Throughput が高くなるわけではなく，Throughput のピークが GPU メモリ使用量の制限下で得られるという予想外の結果となった．

一方，パラメータ数が最も多い Llama3.2-3B では，gpu-memory-utilization が大きいほど Throughput が高くなるという異なる傾向が確認された．Throughput のピークは gpu-memory-utilization: 1.0 の 3.73 infer/sec であり，3 つのモデルの中で最も低くなった．これらのことから，Llama3.2-3B で高スループットを実現するには GPU1 基分のリソース量 (16GB) では不十分である可能性があると考え，使用する GPU を 2 基に増やして再度測定を行うことにした．

2.2.2　GPU2 基使用（モデル: Llama3.2-3B）

Llama3.2-3B で GPU を 2 基 (32GB) 使用して推論を実行した際の Throughput の変化を図 2 に示す．GPU 数が異なるため，同じ指標で比較できるように横軸は実際の GPU メモリ使用量 (GB) としている．GPU を 2 基使用した場合の Throughput ピーク時の GPU メモリ使用量は 21.19 GB となり，GPU1 基あたりのメモリ量である 16GB を上回っていた．この値に到達するまでは，GPU メモリ使用量が多いほど Throughput が

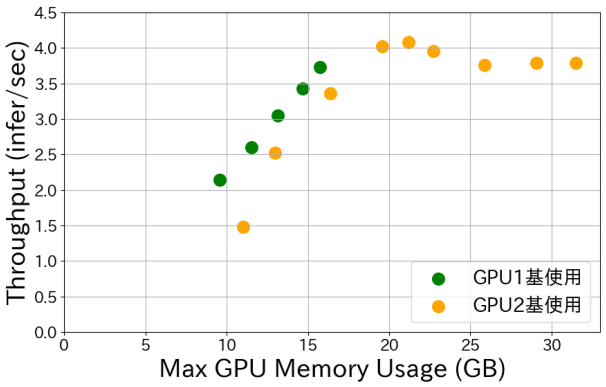

図 2: GPU 数を増やした場合の Throughput の比較 (モデル: Llama3.2-3B)

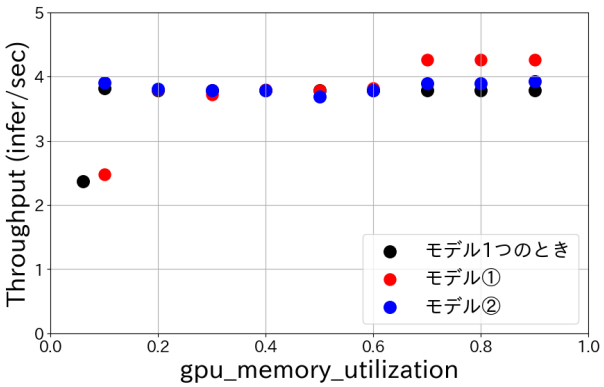

図 3: モデル数を増やした場合の Throughput の比較 (モデル: gpt2)

高くなる傾向が見られたが，その値を超えるとそれ以上 GPU メモリ使用量を増やしても Throughput が高くなることはなかった．以上のことから Llama3.2-3B を実行する際に高い Throughput を引き出すために最低限必要なメモリ量は 21.19 GB であるということが明らかとなった．

2.2.3 複数モデルの同時実行（モデル: gpt2）

gpt2 では，gpu-memory-utilization: 0.1 以上であれば GPU メモリを制限することによる性能の低下が見られないことが明らかとなったため，続いて同一のモデル (gpt2) を複数ロードして複数の推論を同時に実行した際の性能評価を行った．なお，以降の実験では GPU を 1 基のみ使用している．

gpt2 を 2 つ Triton Inference Server にロードして，それぞれで同時に推論を実行した結果を図 3 に示す．その際，gpu-memory-utilization は 2 つのモデルでいずれも同じ値に設定している．横軸は gpu-memory-utilization で，縦軸は Throughput (infer/sec) となっており，モデルを 2 つ同時に動作させたときの各モデルの Throughput をモデル①・モデル②として記載している．比較対象として，モデルを単独で使用していた際の Throughput もプロットしている．実験の結果，gpu-memory-utilization: 0.1 のときはモデル①の Throughput が低くなったが, gpu-memory-utilization: 0.2 以上では，同時に動作させるモデル数を 2 つに増やすことによる Throughput の低下は見られないということがわかった．特に gpu-memory-utilization: 0.7 以上については，モデルを 2 つ同時に動作させたときの方が高い Throughput を記録した．

このときの GPU 使用率を図 4 に示す．モデルを 2 つに増やすことによって，gpu-memory-utilization の値に関わらず GPU 使用率が 20-30%高まっていることがわかる．以上より, 同時に動作させるモデル数を増やすことによって，GPU 使用率を高め，かつ Throughput の低下を引き起こさないということが可能であることが明らかとなった．

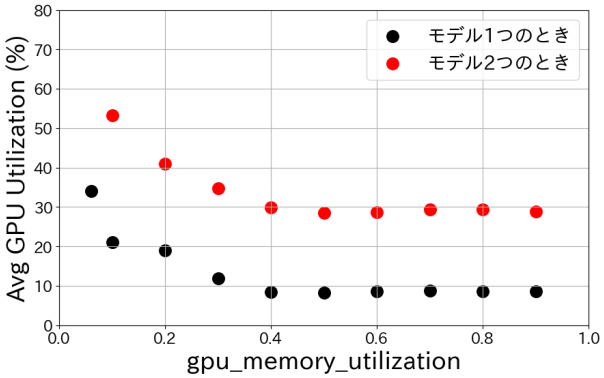

図 4: GPU 使用率の比較 (モデル: gpt2)

3 まとめと今後の課題

本研究では，3 つのモデル (gpt2・Llama3.2-1B・Llama3.2-3B) を用いて GPU リソースを制限して LLM 推論を実行した際の性能評価を行った．その結果，モデルごとに GPU リソースを制限することによって異なる影響が生じ，モデルによってはリソース制限による性能低下の影響が小さいということが明らかとなった．特に gpt2 では，同時に動作させるモデル数を増やすことによって，Throughput の低下を引き起こさずに GPU 使用率を高めることが可能であるということがわかった．さらに gpt2 や Llama3.2-1B，GPU2 基使用時の Llama3.2-3B のようにモデルに対して GPU リソースが十分にある状況では，GPU メモリ使用量に制限を設けた方が Throughput が高くなるという現象が見られた．

今後はその理由を明らかにするため, 異なるパラメータ設定や他の計測ツールを使用するなどして追加調査を行っていく．

参考文献

[1] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Trans. Knowl. Discov. Data*, 18(6), April 2024.

[2] Engine Arguments. `https://docs.vllm.ai/en/v0.4.2/models/engine_args.html`.

[3] Performance Analyzer. `https://docs.nvidia.com/deeplearning/triton-inference-server/archives/triton-inference-server-2310/user-guide/docs/user_guide/perf_analyzer.html`.