



Benchmark em IA

Rafael Jun Teramae Dantas

Universidade De São Paulo

Instituto de Ciências Matemáticas e de Computação

Bacharelado em Sistemas de Informação

SSC0510 - Arquitetura de Computadores - 2º Semestre 2022

Sumário

1. Introdução	2
1.1. Motivação	2
1.2. Sobre o tema	3
2. Desenvolvimento	3
2.1. TOPS	3
2.2. MLPerf	3
2.3. Dificuldade e comparação com benchmark de GPU	4
3. Conclusão	6
4. Referências	8

1. Introdução

1.1. Motivação

A inteligência artificial (IA) é um tema muito comum nas histórias de ficção científica, mas uma de suas primeiras aparições é no famoso romance Frankenstein[1] escrito em 1818 pela autora inglesa Mary Shelley[2]. Este romance conta a história de um cientista que cria uma criatura a partir de restos mortais e eletricidade. Sem entender esse novo mundo, essa criatura vive na marginalidade da sociedade observando essa nova realidade e tentando aprender com o que observava.

Entretanto, um dos melhores e mais famosos exemplos de IA na ficção é nos filmes de Star Wars[3], onde é possível encontrar robôs autônomos, capazes de realizar tudo e mais um pouco do que um ser humano é capaz de fazer.

Carros autônomos[4], robôs que são capazes de conversar e interagir com um locutor[5], computadores que são capazes de derrotar campeões em jogos competitivos (xadrez[6], dota2[7]) e IA que são capazes de criar obras de artes e ganhar competições internacionais[8] são apenas alguns exemplos da inteligência artificial na atualidade. Porém, para alcançar esses produtos finais, são necessários muitos testes.



"Théâtre D'opéra Spatial"

¹ Obra de arte criada pela IA Midjourney que ganhou o Colorado State Fair's annual art competition

1.2. Sobre o tema

Inteligência artificial[9] em computação é uma área que busca desenvolver sistemas capazes de realizar tarefas ou problemas que, na maioria das vezes, somente humanos são capazes de resolver, como por exemplo dirigir um carro ou reconhecer alguém por uma foto.

Benchmark[10] é uma prática adotada por empresas para medir a performance de um certo produto usando certos métodos e certas medidas. Muitas vezes cria-se consórcios para criar testes ou tarefas que serão usadas como teste, com o objetivo de padronizar essas métricas.

Benchmark em IA é medir a performance dos aceleradores usados em aplicações de IA. Diferente de um benchmark padrão, por exemplo de uma placa de vídeo, há diversos testes diferentes para avaliar a performance de um mesmo acelerador, por esse motivo, é complicado avaliar a real performance. Uma solução adotada pela indústria foi criar uma medida (TOPS e suas variantes) e um consórcio para padronizar essa tarefa de avaliação.

2. Desenvolvimento

Antes de entender como o benchmark em IA é feito, é necessário entender primeiro as medidas/métricas que são utilizadas para avaliar o desempenho.

2.1. TOPS

TOPS (Tera Operations per Second)[12][13][14] é uma medida abstrata e simplificada criada para medir a quantidade de operações que um acelerador, com 100% de utilização, consegue calcular em um segundo. Essas operações podem ser entre números de diferentes formatos[11], os mais utilizados são: INT8 (inteiro 8 bits), INT16 (inteiro 16 bits), FP16 (float 16 bits), FP32 (float 32 bits) e FP64 (float 64 bits)².

2.2. MLPerf

MLPerf[17] é um consórcio que reúne o meio acadêmico e a indústria com o objetivo de criar testes justos e imparciais para serem usados na avaliação de benchmark. Esse consórcio tem o objetivo de acelerar o desenvolvimento na área de IA.

² Nos casos que utilizam float pointer (FP), ao invés de utilizar TOPS, utiliza-se TFLOPS (Tera Floating Operations Per Second[18])

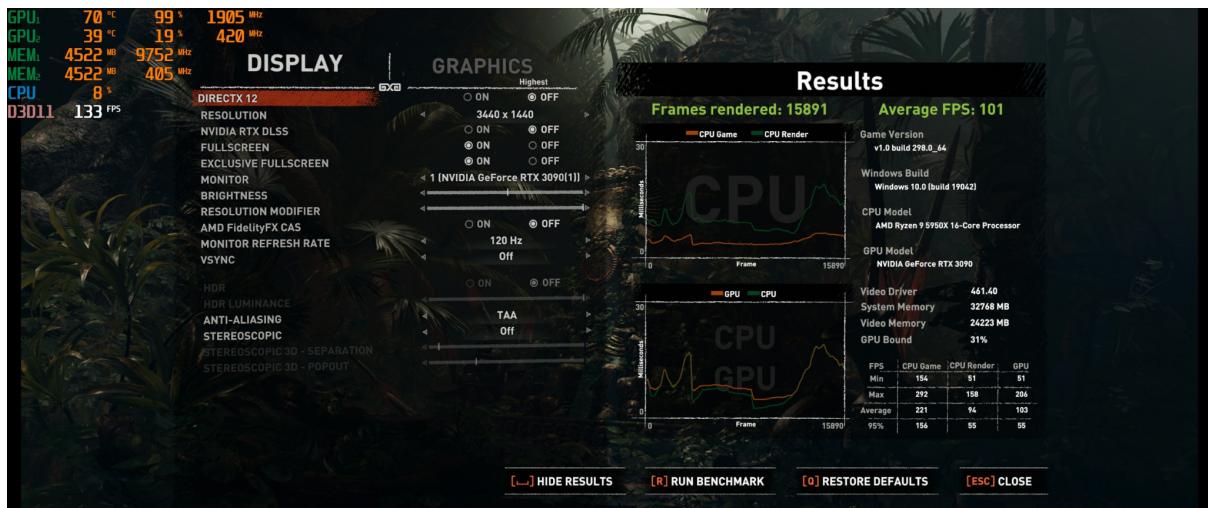
Atualmente, esse consórcio possui 11 testes de diferentes áreas[15][16]:

- Classificação de imagens;
- Detecção de objetos leves;
- Detecção de objetos pesados;
- Segmentação de imagem em biomedicina;
- Reconhecimento automático de fala (ASR);
- Processamento de linguagem natural (NLP);
- Recomendação;
- Aprendizado por reforço;
- Climate Atmospheric River Identification;
- Cosmology Parameter Prediction;
- Quantum Molecular Modeling.

2.3. Dificuldade e comparação com benchmark de GPU

Agora que conhecemos bem resumidamente as medidas/métricas/testes usados na avaliação de benchmark em IA, é possível entender um pouco da dificuldade desse processo. Como na área de IA existem várias aplicações diferentes, é muito difícil centralizar e criar um método de benchmark universal para IA.

Se compararmos o benchmark em IA com o Benchmark de GPUs, é muito mais simplificado e fácil realizar um benchmark em uma GPU. Hoje em dia, um dos métodos mais comuns de benchmark em GPU é testar elas em jogos (principal mercado das GPUs). Muitos jogos possuem opções de benchmark automático já implementadas, facilitando ainda mais o processo[19].



3

³ Resultado da benchmark do jogo Shadow of the Tomb Raider utilizando uma NVIDIA GeForce RTX 3090

Após ativar a opção de benchmark, é possível verificar a temperatura da GPU, porcentagem de utilização, utilização da memória, quantos frames que estão rodando, entre outras informações.

	A100 80GB PCIe	A100 80GB SXM
FP64		9.7 TFLOPS
FP64 Tensor Core		19.5 TFLOPS
FP32		19.5 TFLOPS
Tensor Float 32 (TF32)		156 TFLOPS 312 TFLOPS*
BFLOAT16 Tensor Core		312 TFLOPS 624 TFLOPS*
FP16 Tensor Core		312 TFLOPS 624 TFLOPS*
INT8 Tensor Core		624 TOPS 1248 TOPS*

4

Comparando com uma benchmark em IA, para simplificar, somente as informações da quantidade de TOPS obtidos utilizando certo formato de dado, geralmente, são importantes. Entretanto, caso queira saber da utilização de um acelerador em algum caso específico na área de IA, é preciso recorrer aos testes do MLPerf.

#	Accelerator	#	Software	Benchmark results (minutes)							
				Image classification	Image segmentation (medical)	Object detection, light-weight	Object detection, heavy-weight	Speech recognition	NLP	Recommendation	Reinforcement Learning
				ImageNet	KITS19	OpenImages	COCO	LibriSpeech	Wikipedia	1TB Clickthrough	Go
2	NVIDIA A100-PCIe-80GB	4	MxNet NVIDIA Release 22.04	59.965	58.846						
2	NVIDIA A100-PCIe-80GB	4	PyTorch NVIDIA Release 22.04			217.008		64.106			
2	NVIDIA A100-PCIe-80GB	4	PyTorch NVIDIA Release 22.09				84.853			46.257	
2	NVIDIA A100-SXM4-80GB	8	merlin_hugect NVIDIA Release 22.04								1.849
2	NVIDIA A100-SXM4-80GB	8	MxNet NVIDIA Release 22.04	29.503	24.116						
2	NVIDIA A100-SXM4-80GB	8	PyTorch NVIDIA Release 22.04			87.093		33.990			
2	NVIDIA A100-SXM4-80GB	8	PyTorch NVIDIA Release 22.09				41.804				
8	NVIDIA A100-SXM4-80GB	32	PyTorch NVIDIA Release 22.04				13.675				
16	NVIDIA A100-SXM4-80GB	64	MxNet NVIDIA Release 22.04	4.607							
18	NVIDIA A100-SXM4-80GB	72	MxNet NVIDIA Release 22.04		3.437						
32	NVIDIA A100-SXM4-80GB	128	PyTorch NVIDIA Release 22.04					17.380			
2	NVIDIA A100-SXM4-80GB	8	PyTorch NVIDIA Release 22.09						16.828		
16	NVIDIA A100-SXM4-80GB	64	PyTorch NVIDIA Release 22.09						2.698		

5

⁴ Informações de TOPS de uma NVIDIA A100[20]

⁵ Alguns resultados de testes do MLPerf em uma NVIDIA A100[21]

3. Conclusão

O mercado de inteligência artificial é um mercado que vem crescendo cada vez mais, seja por causa das possibilidades quase infinitas que a IA traz ou seja por causa das necessidades da nossa realidade de automatizar tarefas. Um grande fator que possibilitou esse crescimento foi a evolução dos hardwares, diminuindo o custo por GFLOPS[22].

Date	Approximate USD per GFLOPS		Platform providing the lowest cost per GFLOPS
	Unadjusted	2021 ^[69]	
1945	\$129.49 trillion	\$1.88 quadrillion	ENIAC: \$487,000 in 1945 and \$7,195,000 in 2019
1961	\$18.7 billion	\$169.6 billion	A basic installation of IBM 7030 Stretch had a cost at the time of US\$7.78 million each.
1984	\$18,750,000	\$48,900,000	Cray X-MP/48
1997	\$30,000	\$51,000	Two 16-processor Beowulf clusters with Pentium Pro microprocessors ^[71]
April 2000	\$1,000	\$1,600	Bunyip Beowulf cluster
May 2000	\$640	\$1,021	KLAT2
August 2003	\$82	\$121	KASYO
August 2007	\$48	\$63	Microwulf
March 2011	\$1.80	\$2.19	HPU4Science
August 2012	\$0.75	\$0.89	Quad AMD Radeon 7970 System
June 2013	\$0.22	\$0.26	Sony PlayStation 4
November 2013	\$0.16	\$0.19	AMD Sempron 145 & GeForce GTX 760 system
December 2013	\$0.12	\$0.14	Pentium G550 & Radeon R9 290 system
January 2015	\$0.08	\$0.09	Celeron G1830 & Radeon R9 295X2 system
June 2017	\$0.06	\$0.07	AMD Ryzen 7 1700 & AMD Radeon Vega Frontier Edition system

October 2017	\$0.03	\$0.03	Intel Celeron G3930 & AMD RX Vega 64 system
November 2020	\$0.03	\$0.03	AMD Ryzen 3600 & 3x NVIDIA RTX 3080 system
November 2020	\$0.04	\$0.04	PlayStation 5
November 2020	\$0.04	\$0.04	Xbox Series X
September 2022	\$0.02	\$0.02	RTX 4090

6

Entretanto, não basta apenas desenvolver novas tecnologias e não testá-las, por isso é muito importante medidas/métricas e consórcios com o objetivo de generalizar e acelerar os processos de benchmark, visto as dificuldades e as diferentes aplicabilidades da inteligência artificial nos dias de hoje.

⁶ Tabela mostrando a plataforma com o menor custo por GFLOPS por ano

4. Referências

1. [Frankenstein - Wikipedia](#)
2. [Mary Shelley - Wikipedia](#)
3. [Star Wars - Wikipedia](#)
4. [Full Self-Driving](#)
5. [Ameca - Engineered Arts](#)
6. [Deep Blue vs Kasparov: How a computer beat best chess player in the world - BBC News](#)
7. [OpenAI Five Beats World Champion DOTA2 Team 2-0! 🤖](#)
8. [AI-Generated Art Won a Prize. Artists Aren't Happy. - The New York Times](#)
9. [Benchmarking - Wikipedia](#)
10. [Artificial intelligence - Wikipedia](#)
11. [FP64, FP32, FP16, BFLOAT16, TF32, and other members of the ZOO | by Grigory Sapunov](#)
12. [TOPS vs. real world performance: Benchmarking performance for AI accelerators - Embedded.com](#)
13. [As AI chips improve, is TOPS the best way to measure their power? | VentureBeat](#)
14. [TOPS, Memory, Throughput And Inference Efficiency](#)
15. [Benchmarks de AI MLPerf | NVIDIA](#)
16. [MLPerf AI Benchmarks | NVIDIA](#)
17. [Philosophy | MLCommons](#)
18. [FLOPS – Wikipédia, a encyclopédia livre](#)
19. [How to benchmark your graphics card | PCWorld](#)
20. [NVIDIA A100](#)
21. [v2.1 Results | MLCommons](#)
22. [FLOPS - Wikipedia](#)