Genome GPT

Member:
Ke Ding, ANU

Mentors:
Hyungon Ryu, Nvidia
Zhuochen Wu, NCI

Genome GPT

- Problem trying to solve
  - Building a DNA sequence base model for various downstream tasks: expression, binding sites
- Scientific driver for the chosen algorithm
  - Transformer based NNs have shown great potential
- What's the algorithmic motif?
- What parts are you focusing on?
  - Modifying Transformer for long DNA sequence
  - Running Transformer on multiple nodes

Evolution and Strategy

- What was your goal coming here?
  - Running GPT models on multi nodes and speeding up by at least 5 times
- What was your initial strategy?
  - Using the Horovod + Tensorflow framework
- How did this strategy change?
  - Using the DeepSpeed + PyTorch framework

Results and Final Profile

- What were you able to accomplish?
  - Did you achieve speed up?

    (*show multi-core CPU vs GPU numbers)

    Yes. 24 times from cpu to gpu

- What did you learn?
  - Using DeepSpeed to scale up the NN training
  - Using Nsight System to profile the training
  - A bigger picture how things work

Energy Efficiency

The calculator will compare energy consumption of a number of CPU only nodes with dual CPUs required to perform the same amount of work as 1 GPU node with 2 CPUs and 8 GPUs.

INPUTS	
# CPU Cores	12
# GPUs (A100)	1
Application Speedup	22.4x
Node Replacement	16.8x

GPU NODE POWER SAVINGS			
	AMD Dual Rome 7742	8x A100 80GB SXM4	Power Savings
Compute Power (W)	18,480	6,500	11,980
Networking Power (W)	780	93	687
Total Power (W)	**19,260**	**6,593**	12,667
Node Power efficiency			**2.9x**

ANNUAL ENERGY SAVINGS PER GPU NODE			
	AMD Dual Rome 7742	8x A100 80GB SXM4	Power Savings
Compute Power (kWh/year)	161,885	56,940	104,945
Networking Power (kWh/year)	6,834	814	6,020
Total Power (kWh/year)	**168,719**	**57,754**	**110,965**

$/kWh	$ 0.34
Annual Cost Savings	$ 37,728.18
3-year Cost Savings	$ 113,184.53

Metric Tons of CO2	79
Gasoline Cars Driven for 1 year	17
Seedlings Trees grown for 10 years	1,301

(source: Link)

What problems have you encountered?

- Problems with legacy app structure
  - Changed the code base and encountered errors with DeepSpeed. And we are still working on it.

Wishlist

- What do you wish existed to make your life easier?
  - I got what I need
  - If I have to name one thing… A dedicate DGX A100 in the near future

Was it worth it?

- Was this worth it?
  - Yes
- Will you continue development?
  - Yes
- What sustained resources/support will be critical for your work after the event?
  - Connection with my Mentor
  - Access to gpu clusters(ideally A100)

Application Background

- A pretrain LLM on DNA sequences will benefit the community considering the great success of GPT on natural language
- Genome sequence is very similar to our language (DNA -> Protein -> Human : Character -> Vocabulary -> Language)

```
iter_dt 40.28ms; iter 10: train loss 3.00477; perplexity 20.18150;  bpc 3.00477
iter_dt 40.32ms; iter 20: train loss 2.73704; perplexity 15.44120;  bpc 2.73704
iter_dt 40.26ms; iter 30: train loss 2.63288; perplexity 13.91381;  bpc 2.63288
iter_dt 39.99ms; iter 40: train loss 2.56995; perplexity 13.06516;  bpc 2.56995
iter_dt 40.30ms; iter 50: train loss 2.53902; perplexity 12.66729;  bpc 2.53902
iter_dt 40.16ms; iter 60: train loss 2.50262; perplexity 12.21445;  bpc 2.50262
iter_dt 40.51ms; iter 70: train loss 2.49877; perplexity 12.16757;  bpc 2.49877
iter_dt 40.51ms; iter 80: train loss 2.46701; perplexity 11.78712;  bpc 2.46701
iter_dt 40.49ms; iter 90: train loss 2.45316; perplexity 11.62504;  bpc 2.45316
```

It is working

```
  [turn off auto-refresh]

gadi-gpu-v100-0155.gadi.nci.org.au  Wed Jun  7 17:31:34 2023  525.60.13
[0] Tesla V100-SXM2-32GB       | 46°C,  96 % |  7564 / 32768 MB | kd1348(7294M)
[1] Tesla V100-SXM2-32GB       | 43°C,  93 % |  7556 / 32768 MB | kd1348(7286M)
[2] Tesla V100-SXM2-32GB       | 45°C,  97 % |  7524 / 32768 MB | kd1348(7254M)
[3] Tesla V100-SXM2-32GB       | 47°C,  95 % |  7564 / 32768 MB | kd1348(7294M)
gadi-gpu-v100-0159.gadi.nci.org.au  Wed Jun  7 17:31:35 2023  525.60.13
[0] Tesla V100-SXM2-32GB       | 46°C,  98 % |  7564 / 32768 MB | kd1348(7294M)
[1] Tesla V100-SXM2-32GB       | 43°C,  93 % |  7556 / 32768 MB | kd1348(7286M)
[2] Tesla V100-SXM2-32GB       | 43°C,  98 % |  7524 / 32768 MB | kd1348(7254M)
[3] Tesla V100-SXM2-32GB       | 47°C,  92 % |  7564 / 32768 MB | kd1348(7294M)
```

It is scaling up

Hackathon Objectives and Approach

- Programming models: GPT
- Profiling / hot spots: Nsight
- Libraries: DeepSpeed
- Performance tuning: Nsight

Technical Accomplishments and Impact

- What were you able to achieve at the hackathon?: We build a proof of concept
- How did you achieve it?: With the assist from mentor and the comfortable coding environment
- Speedup: 24 times
- Why does it matter / what does it enable?

Please use 100 words to summarize your team's achievements during this Hackathon

The most valuable thing we achieved is the 'future work' on this project. My mentor Hyungon shares a lot of useful tips and points out several potential improvements I can adopt on this project. He also shares his experience on AlphaFold.

PROMOTING YOUR WORK: AVAILABLE OPPORTUNITIES

- **Papers and Talks:** Please acknowledge the Open Hackathons program and OpenACC Organization in any planned or upcoming papers, presentations, or talks.

  "This work was completed in part at the [Event name], part of the Open Hackathons program. The authors would like to acknowledge OpenACC-Standard.org for their support."
- **Social Media Support:** Please feel free to promote your participation across your social media channels. Tag **@OpenACCorg** and **#OpenHackathons** and we are happy to amplify.
- **Blogs and Technical Write-ups**: Create a blog post or technical article that highlights the work being done and results achieved.
- **Quotes and Testimonials:** Highlight your quote or feedback on our channels (i.e. social, website, etc.).

***Please reach out to Izumi Barker (ibarker@nvidia.com) to discuss marketing options and opportunities.**