TRAINING MULTI-BILLION PARAMETER LANGUAGE MODELS USING MEGATRON

Mohammad Shoeybi
Hot Chips Conf., Aug. 16, 2020
MODEL SIZE TREND IN NLP

- Training the largest transformer-based language model has recently been one of the best ways to advance the state-of-the-art in NLP applications.

- NLP model size increases by almost an order of magnitude every year.
WHY LARGE NLP MODELS?

- Imagenet moment of NLP
- Unsupervised pretraining on large text corpora has eliminated training dataset size issues
- Lots of downstream NLP applications have benefited from recent advancements
- Training larger models with more data results in better accuracy in almost all cases
EXAMPLE 1: LEFT-TO-RIGHT LANGUAGE MODEL

- GPT2-based language model
- Training data: 174 GB WebText / CC-Stories / Wikipedia / RealNews

<table>
<thead>
<tr>
<th>Model Size</th>
<th>Wikitext-103 (Perplexity ↓)</th>
</tr>
</thead>
<tbody>
<tr>
<td>355 M</td>
<td>19.22</td>
</tr>
<tr>
<td>2.5 B</td>
<td>12.68</td>
</tr>
<tr>
<td>8.3 B</td>
<td>10.81</td>
</tr>
<tr>
<td>Previous SOTA</td>
<td>16.43*</td>
</tr>
</tbody>
</table>

* Dynamic Evaluation of Transformer Language Models, Krause et. al., 2019
Context: "You have stolen my dreams and my childhood with your empty words and yet, I'm one of the lucky ones," Thunberg told the assembly, her signature long braid swept to the side. "People are suffering, people are dying."

Megatron-LM:
POWERFUL GENERATIONS

Context: “You have stolen my dreams and my childhood with your empty words and yet, I’m one of the lucky ones,” Thunberg told the assembly, her signature long braid swept to the side. “People are suffering, people are dying.”

Megatron-LM:

In Australia, a similar push has been led by 16-year-old climate activist Greta Thunberg, who has led protests in the Swedish capital of Stockholm.

On Saturday, Thunberg joined forces with the well-known climate activist and former U.S. vice president Al Gore to lead a “Live Earth” concert in the Norwegian capital of Oslo.

The concert and protest was aimed at drawing attention to the need to reduce carbon emissions.

Thunberg told the crowd that even though she was tired, “I am happy that I am here. I am happy that I am not just doing it for myself, but that I am doing it for all of you.”

Thunberg's activism has also led her to become a target of those who deny climate change is an issue.

In September, a group of Swedish men made death threats against her on a Facebook page, according to Swedish media. The threats came after
POWERFUL GENERATIONS

Context: "You have stolen my dreams and my childhood with your empty words and yet, I'm one of the lucky ones." Thunberg told the assembly, her signature long braid swept to the side. "People are suffering, people are dying."

Megatron-LM:
In Australia, a similar push has been led by 16-year-old climate activist Greta Thunberg, who has led protests in the Swedish capital of Stockholm.

On Saturday, Thunberg joined forces with the well-known climate activist and former U.S. vice president Al Gore to lead a "Live Earth" concert in the Norwegian capital of Oslo.

The concert and protest was aimed at drawing attention to the need to reduce carbon emissions.

Thunberg told the crowd that even though she was tired, "I am happy that I am here. I am happy that I am not just doing it for myself, but that I am doing it for all of you."

Thunberg's activism has also led her to become a target of those who deny climate change is an issue.

In September, a group of Swedish men made death threats against her on a Facebook page, according to Swedish media. The threats came after she spoke at a conference in New York.

"I don't care if they send me death threats," she told Swedish radio at the time. "I have been bullied by the media before and it's not going to change me."

Thunberg has vowed to continue her activism, and she has called on young people across the globe to lead the fight against climate change.
EXAMPLE 2: BERT

- Canonical downstream tasks
- Larger model trained on fewer number of tokens produces better results

<table>
<thead>
<tr>
<th>Model</th>
<th>Trained tokens (ratio)</th>
<th>MNLI\textsuperscript{t} m/mm accuracy</th>
<th>QQP\textsuperscript{t} accuracy</th>
<th>SQuAD 1.1\textsuperscript{t} F1/EM</th>
<th>SQuAD 2.0\textsuperscript{t} F1/EM</th>
<th>RACE m/h\textsuperscript{*} accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>RoBERTa</td>
<td>2</td>
<td>90.2 / 90.2</td>
<td>92.2</td>
<td>94.6 / 88.9</td>
<td>89.4 / 86.5</td>
<td>86.5 / 81.3</td>
</tr>
<tr>
<td>ALBERT</td>
<td>3</td>
<td>90.8</td>
<td>92.2</td>
<td>94.8 / 89.3</td>
<td>90.2 / 87.4</td>
<td>89.0 / 85.5</td>
</tr>
<tr>
<td>XLNet</td>
<td>2</td>
<td>90.8 / 90.8</td>
<td>92.3</td>
<td>95.1 / 89.7</td>
<td>90.6 / 87.9</td>
<td>88.6 / 84.0</td>
</tr>
<tr>
<td>Megatron-334M</td>
<td>1</td>
<td>89.7 / 90.0</td>
<td>92.3</td>
<td>94.2 / 88.0</td>
<td>88.1 / 84.8</td>
<td>86.9 / 81.5</td>
</tr>
<tr>
<td>Megatron-1.3B</td>
<td>1</td>
<td>90.9 / 91.0</td>
<td>92.6</td>
<td>94.9 / 89.1</td>
<td>90.2 / 87.1</td>
<td>90.4 / 86.1</td>
</tr>
<tr>
<td>Megatron-3.9B</td>
<td>1</td>
<td>91.4 / 91.4</td>
<td>92.7</td>
<td>95.5 / 90.0</td>
<td>91.2 / 88.5</td>
<td>91.8 / 88.6</td>
</tr>
</tbody>
</table>
**EXAMPLE 3: TEACHING MODELS TO GENERATE QUESTIONS AND ANSWERS**

1. Generate Text
   \[ \hat{c} \sim p(c) \]

   - Using synthetic data can beat using only human labeled data

<table>
<thead>
<tr>
<th>Text Source</th>
<th>Data Size</th>
<th>finetune data</th>
<th># Questions</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wikipedia</td>
<td>638 MB</td>
<td>Synthetic</td>
<td>19,925,130</td>
<td>88.4</td>
<td>94.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+SQUAD</td>
<td>20,012,729</td>
<td><strong>89.4</strong></td>
<td><strong>95.2</strong></td>
</tr>
<tr>
<td>8.3B GPT-2</td>
<td>480 MB</td>
<td>Synthetic</td>
<td>17,400,016</td>
<td>88.4</td>
<td>93.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+SQUAD</td>
<td>17,487,615</td>
<td><strong>89.1</strong></td>
<td><strong>94.9</strong></td>
</tr>
<tr>
<td>SQUAD 1.1</td>
<td>14MB</td>
<td>SQUAD</td>
<td>87,599</td>
<td>87.7</td>
<td>94.0</td>
</tr>
</tbody>
</table>

   - Scale is the key in improving results

2. Extract Answers From Text
   \[ \hat{a} \sim p(a|\hat{c}) \]

   Context: “I Got Mine” is a song by American rapper 50 Cent from his debut studio album “Get Rich or Die Tryin’” (2003). The song features a guest appearance from fellow New York City rapper Nas, who was also featured on the previous single from “Get Rich or Die Tryin’”, “Hate Me Now”.

3. Pose Questions From Answers
   \[ \hat{q} \sim p(q|\hat{a}, \hat{c}) \]

   - Q: What was the title of 50 Cent’s first Album?
   - Q: What was the title of 50 Cent’s second Album?

4. Filter out bad question answer pairs
   \[ \hat{a} \not\sim a^* \sim p(a|\hat{c}, \hat{q}) \]

EXAMPLE 3: Q&A OVER TESLA CAR MANUAL

Tesla car manual been used to get CONTEXT!!

Please enter your question: What replacement battery does my key fob need?

Short Answer: CR2032

Long Answer: The key’s battery lasts for approximately a year. When the battery is low, a message displays on the instrument panel. Follow these steps to replace it: 1. With the key placed button side down on a soft surface, use a small flat-bladed tool or fingernail to release the bottom cover. 2. Remove the battery by lifting it away from the front retaining clips. 3. Insert the new battery (type CR2032) with the “+” side facing up. Note: CR2032 batteries can be purchased through online retailers, local supermarkets and drug stores. Note: Wipe the battery clean before fitting and avoid touching the battery’s flat surfaces. Finger marks on the flat surfaces of the battery can reduce battery life. 4. Holding the cover at an angle, align the tabs on the widest side of the cover with the corresponding slots on the key, then press the cover firmly onto the key until it snaps into place.

Please enter your question: 
EXAMPLE 4: CHATBOTS

- Generative Conversation Control (GCC)

- Persona control by conditioning on prior conversations

<table>
<thead>
<tr>
<th># Parameters</th>
<th>Data Source</th>
<th>Persona Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>DLGNet</td>
<td>345 Million</td>
<td>Movie Triples / Ubuntu Dialogue Corpus</td>
</tr>
<tr>
<td>DiaLoGPT</td>
<td>768 Million</td>
<td>Reddit (2005-2017)</td>
</tr>
<tr>
<td>Meena</td>
<td>2.6 Billion</td>
<td>Public Social Media</td>
</tr>
<tr>
<td>GCC</td>
<td>8.3 Billion</td>
<td>Reddit (2019)</td>
</tr>
</tbody>
</table>

**Speaker** | **Conversation Turn** |
---|---|
A | They are worried about themes becoming an exploit. It happened multiple times with the *3DS* |
B | How was themes an exploit on the 3ds |
A | You would inject the theme and holding L or R at boot would start the homebrew launcher |
B | *Thanks* I was not aware of that and thought I would learn a new thing |
A | *I mean, 3DS* has been out for some time |
B | *yeah but* I only started playing it in december 2018 |

(i) # (Ref. Parent Comment) → Ref. Reply Comment for Speaker B

1. \((n/a)\) → I once had 100 pings from the same channel and all was because the owner did not know how to make a long comment
2. (You’re worse than us.) → And im one of them but I do got skins im just just as bad
3. \((n/a)\) → Oh well *I am a newbie*
4. \((n/a)\) → oh wow I did not see the read it backwards thing so I did not understand only when I scrolled down

(ii)
EXAMPLE 4: CHATBOTS

======== Conversation ========
[Your Turn]
> Who do you think will win the 2020 elections?

[Bot's Turn]
> a) Trump
> b) Harris
> c) Warren
> d) Klobuchar
> Probably Harris.

[Your Turn]
> Why do you think Harris over the other candidates?

[Bot's Turn]
> she seems more appealing to me.

[Your Turn]
> But what about her is appealing?
RECAP: WHY LARGE NLP MODELS?

- Imagenet moment of NLP
- Unsupervised pretraining on large text corpora has eliminated training dataset size issues
- Lots of downstream NLP application have benefited from recent advancements
- Training larger models with more data results in better accuracy in all cases

We need software and hardware that can efficiently and easily enable researchers to train large NLP models
NVIDIA’s framework for efficiently training the world’s largest transformer-based language models.

Repo: https://github.com/NVIDIA/Megatron-LM

- All the aforementioned cases are trained using Megatron
- Megatron has been directly used to train Turing-NLG (17.2B)
- Megatron has inspired other work such as BlenderBot by Facebook
GOALS FOR MEGATRON

- Low barrier to entry
  - Ability to build on top our existing code base (no rewriting, ...)
  - Devising simple methods that require minimal changes

- Training of transformer-based language models with billions of parameters
  - Requires model parallelism to fit in GPU memory
  - Only support transformer-based model

- Achieving high utilization and scaling up to thousands of GPUs
MODEL PARALLELISM

- **Inter-Layer (Pipeline) Parallelism**
  - Split sets of layers across multiple devices
  - Layer 0, 1, 2 and layer 3, 4, 5 are on different devices

- **Intra-Layer (Tensor) Parallelism**
  - Split individual layers across multiple devices
  - Both devices compute different parts of Layer 0, 1, 2, 3, 4, 5
WHY INTRA-LAYER MODEL PARALLELISM

• Tensor parallelism is much simpler to implement

• Easier to load-balance

• Less restrictive on the batch-size (bubble issue in pipelining)
  • Intra-layer model parallelism is orthogonal to pipeline parallelism: very large models such as GPT-3 use both.

• Transformers have large GEMMs
  • Tensor parallelism works well for large matrices

• NVIDIA DGX-A100 boxes with nvlink
  • DGX-A100 has 600 GB/sec GPU-to-GPU bidirectional bandwidth
CHALLENGES WITH INTRA-LAYER MODEL PARALLELISM

- Tensor splitting results in lower math intensity
  - This approach is not suitable for strong scaling

- We should make sure math intensity stays above that of the processor
  - A100 math intensity = 312 teraFLOPs/1555 GB/sec = 200

- Intra-layer model parallelism requires more communication.
  - DGX-A100 with nvlink mitigates this issue.

### Parallel

**Operation:** \( Y_{n \times (n/p)} = X_{n \times n} A_{n \times (n/p)} \)

**Flops:** \( \frac{2n^3}{p} \)

**Bandwidth:** \( 2n^2 \left(1 + \frac{2}{p}\right) \)

**Intensity:** \( \frac{1}{2 + \frac{n}{p}} \)

SIMPLE EXAMPLE OF TENSOR PARALLELISM

Row Parallel Linear Layer

\[
A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}
\]

\[
X = [X_1, X_2]
\]

**Forward**: \( Y = Y_1 + Y_2 \) (all-reduce)

**Backward**: \( \frac{\partial L}{\partial Y_i} = \frac{\partial L}{\partial Y} \) (identity)

**Function**: \( f : X_1 \) (split)

**Backward**: \( \frac{\partial L}{\partial X} = [\frac{\partial L}{\partial X_1}, \frac{\partial L}{\partial X_2}] \) (all-gather)

Column Parallel Linear Layer

**Forward**: \( Y = [Y_1, Y_2] \) (all-gather)

**Backward**: \( \frac{\partial L}{\partial Y_i} \) (split)

**Function**: \( f : X \) (identity)

**Backward**: \( \frac{\partial L}{\partial X} = \frac{\partial L}{\partial X}_1 + \frac{\partial L}{\partial X}_2 \) (all-reduce)
APPROACH FOR TRANSFORMER MODELS

- Develop an approach that can be fully implemented with insertion of few simple collectives
- Rely on pre-existing NCCL/PyTorch operations for a native PyTorch implementation
- Group math heavy operations (such as GEMMs) before parallel communication point
PARTITIONING MLP

- **MLP:**
  \[
  Y = \text{GeLU}(XA) \\
  Z = \text{Dropout}(YB)
  \]

- **Approach 1:** split X column-wise and A row-wise:
  \[
  X = [X_1, X_2]  \\
  A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}  \\
  Y = \text{GeLU}(X_1 A_1 + X_2 A_2)
  \]
  - Before GeLU, we will need a communication point

- **Approach 2:** split A column-wise:
  \[
  A = [A_1, A_2]  \\
  [Y_1, Y_2] = [\text{GeLU}(X A_1), \text{GeLU}(X A_2)]
  \]
  - no communication is required
MLP

\[ Y = \text{GeLU}(XA) \]

\[ Z = \text{Dropout}(YB) \]

\[ A = [A_1, A_2] \]

\[ B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \]

\( f \) and \( g \) are conjugate, \( f \) is identity operator in the forward pass and all-reduce in the backward pass while \( g \) is all-reduce in forward and identity in backward.
f AND g ARE SIMPLE

class f(torch.autograd.Function):
    def forward(ctx, x):
        return x
    def backward(ctx, gradient):
        all_reduce(gradient)
        return gradient
PARTITIONING: SELF-ATTENTION

- Self-attention is more complex than MLP

Scaled Dot-Product Attention

Multi-Head Attention

Figure courtesy of Vaswani et al. 2017
SELF-ATTENTION

\[ Y = \text{Self-Attention}(X) \]

\[ \begin{align*}
    &X \xrightarrow{f} \{ V_1, Q_1, K_1, \text{Softmax}, \text{Dropout} \} \rightarrow Y_1 \\
    &X \xrightarrow{f} \{ V_2, Q_2, K_2, \text{Softmax}, \text{Dropout} \} \rightarrow Y_2 \\
    &V = [V_1, V_2] \\
    &K = [K_1, K_2] \\
    &Q = [Q_1, Q_2] \end{align*} \]

\[ Z = \text{Dropout}(YB) \]

\[ B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \]

\( f \) and \( g \) are conjugate, \( f \) is identity operator in the forward pass and all-reduce in the backward pass while \( g \) is all-reduce in forward and identity in backward.
PARALLEL TRANSFORMER LAYER

2 All-Reduce (forward + backward)

2 All-Reduce (forward + backward)
HYBRID MODEL + DATA PARALLELISM
WEAK SCALING EFFICIENCY ON SELENE

Baseline (1.2B parameters on a single GPU) sustains **118 teraFLOPs/sec** on an A100 GPU during the entire training process.
SIDE CHALLENGES OF TRAINING LARGE MODELS
CHANGES IN MODEL STRUCTURE AND INITIALIZATION

• Scaling BERT model as presented in the original work beyond 336M parameters results in instabilities

• Rearranging the residual connection to allow for the direct flow of gradients is necessary

• However, the weights right before the residual connection need to be initialized with a much lower variance
RANDOM NUMBERS IN SELF-ATTENTION

- Tensor splitting results in both serial and parallel regions
- For a model instance
  - serial regions need to have the same RNG state
  - parallel regions need to have different RNG states
- As a result we need to track two sets of RNG states
LARGE MODELS ARE MORE SENSITIVE TO DATA SHUFFLING

- Small models are insensitive to the dataset order
- Larger models have much higher power of memorization and as a result more sensitive to shuffling scheme
- Special attention to dataloaders is required for large models
SUMMARY

- Ability to train large language models is essential for today’s NLP application.
- Intra-layer model parallelism is powerful yet simple for NLP models.
- We can scale models efficiently to billions of parameters on thousands of GPUs.
- When scaling models, careful attention to model structure and data loaders is required.