Scaling up models and data

Colin Raffel

CIFAR Deep Learning & Reinforcement Learning Summer School
From “Real-Time Social Media Analytics with Deep Transformer Language Models: A Big Data Approach” by Ahmet and Abdullah
From “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks” by Tan and Le
From “Deep Learning Scaling is Predictable, Empirically” by Hestness et al.
From “Big Self-Supervised Models are Strong Semi-Supervised Learners” by Chen et al.
From “Scaling Laws for Neural Language Models” by Kaplan et al.
From “Scaling Laws for Neural Language Models” by Kaplan et al.
\[ \hat{y}_i = f_\theta(x_i) \]
\[
\underbrace{... \text{Wh} ...}_{N}
\]
\[ \partial \theta = \sum_{i=1}^{N} \nabla_\theta \mathcal{L}(\hat{y}_i, y_i) \]
\[ \theta \leftarrow \theta + \text{optimizer}(\partial \theta) \]
\[ \hat{y}_i = f_{\theta}(x_i) \]

\[ \cdots W h \cdots \]

\[ \partial \theta = \sum_{i=1}^{N} \nabla_{\theta} \mathcal{L}(\hat{y}_i, y_i) \]

\[ \theta \leftarrow \theta + \text{optimizer}(\partial \theta) \]
\hat{y}_i = f_\theta(x_i) \\
\text{...} \ W \ h \ ... \\
\partial \theta = \sum_{i=1}^{N} \nabla_\theta \mathcal{L}(\hat{y}_i, y_i) \\
\theta \leftarrow \theta + \text{optimizer}(\partial \theta)
$x_1, \ldots, x_N$ → $f_\theta(x)$ → $\sum_{i=1}^{N} \nabla_\theta \mathcal{L}(\hat{y}_i, y_i)$ → $\partial \theta$
Device 1

\[ x_1, \ldots, x_{\frac{N}{2}} \rightarrow f_\theta(x) \rightarrow \sum_{i=1}^{\frac{N}{2}} \nabla_\theta \mathcal{L}(\hat{y}_i, y_i) \rightarrow \partial \theta \]

Device 2

\[ x_{\frac{N}{2}+1}, \ldots, x_N \rightarrow f_\theta(x) \rightarrow \sum_{i=\frac{N}{2}+1}^{N} \nabla_\theta \mathcal{L}(\hat{y}_i, y_i) \]
Device 1

Device 2

concat
From “PipeDream: Generalized Pipeline Parallelism for DNN Training” by Narayanan et al.
From “PipeDream: Generalized Pipeline Parallelism for DNN Training” by Narayanan et al.
From “Efficient Large-Scale Language Model Training on GPU Clusters” by Narayanan et al.
**float32**
- range: \(\sim 1e^{-38} \text{ to } 3e^{38}\)
- 8 bits for the exponent
- 23 bits for the fraction

**float16**
- range: \(\sim 5.9e^{-8} \text{ to } 6.5e^{4}\)
- 5 bits for the exponent
- 10 bits for the fraction

**bfloat16**
- range: \(\sim 1e^{-38} \text{ to } 3e^{38}\)
- 8 bits for the exponent
- 7 bits for the fraction

From [https://cloud.google.com/tpu/docs/bfloat16](https://cloud.google.com/tpu/docs/bfloat16)
From “Mixed Precision Training” by Micikevicius et al.
Forward pass →

← Backward pass
From “Scaling Laws for Neural Language Models” by Kaplan et al.
\[ \mathcal{L} = - \sum_y p_\theta(y|x) \log p_\theta(y|x) \]
\[ \mathcal{L} = -\arg\max_y \left[ p_\theta(y|x) \right] \log p_\theta(y|x) \]
\[ \mathcal{L} = \text{sg} \left[ p_{\theta}(y \mid x) \right] \log p_{\theta}(y \mid x') \]
The cabs charged, used, initially, even, future, became, the, yellow, reported, that, luxurious, horse-drawn, were that, internal, conspicuous, cab the same rates as those charged, used, initially, even, future, became, the, yellow, reported, that, luxurious, horse-drawn, were that, internal, conspicuous, cab by horse-drawn cabs and were quite popular, the Prince of Wales (the King Edward VII) travelled in quite popular, the Prince of Wales (the King Edward VII) travelled in. The cabs quickly became known as "hummingbirds" for noise made by their motors and their distinctive black and livery. Passengers became known as "hummingbirds" for noise made by their motors and their distinctive black and livery. Passengers the interior fittings were conspicuous when compared to cabs but there some complaints the lighting made them too conspicuous to those outside conspicuous.

This movie is terrible! The acting is bad and I was bored the entire time. There was no plot and nothing interesting happened. I was really surprised since I had very high expectations. I want 103 minutes of my life back!

negative
The size of data available on the web has enabled deep learning models to achieve high accuracy on specific benchmarks in NLP and computer vision applications. However, in both application areas, the training data has been shown to have problematic characteristics resulting in models that encode stereotypical and derogatory associations along gender, race, ethnicity, and disability status.

From “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” by Bender et al.
From “Documenting the English Colossal Clean Crawled Corpus” by Dodge et al.
... accuracy improvements depend on the availability of exceptionally large computational resources that necessitate similarly substantial energy consumption. As a result these models are costly to train and develop, both financially, due to the cost of hardware and electricity or cloud compute time, and environmentally, due to the carbon footprint required to fuel modern tensor processing hardware.

From “Energy and Policy Considerations for Deep Learning in NLP” by Strubell et al.
<table>
<thead>
<tr>
<th>Model</th>
<th>Evolved Transformer NAS</th>
<th>T5</th>
<th>Meena</th>
<th>Gshard -600B</th>
<th>Switch Transformer</th>
<th>GPT-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Parameters (B)</td>
<td>0.064 per model</td>
<td>11</td>
<td>2.6</td>
<td>619</td>
<td>1500</td>
<td>175</td>
</tr>
<tr>
<td>Percent of model activated on every token</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>0.25%</td>
<td>0.10%</td>
<td>100%</td>
</tr>
<tr>
<td>Developer</td>
<td>Google</td>
<td>Google</td>
<td>OpenAI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Datacenter of original experiment</td>
<td>Google Georgia</td>
<td>Google Taiwan</td>
<td>Google Georgia</td>
<td>Google North Carolina</td>
<td>Google Georgia</td>
<td>Microsoft</td>
</tr>
<tr>
<td>Datacenter Gross CO₂e/KWh (kg/KWh when it was run)</td>
<td>0.431</td>
<td>0.545</td>
<td>0.415</td>
<td>0.201</td>
<td>0.403</td>
<td>0.429</td>
</tr>
<tr>
<td>Datacenter Net CO₂e/KWh (kg/KWh when it was run)</td>
<td>0.431</td>
<td>0.545</td>
<td>0.415</td>
<td>0.177</td>
<td>0.330</td>
<td>0.429</td>
</tr>
<tr>
<td>Datacenter PUE (when it was run)</td>
<td>1.10</td>
<td>1.12</td>
<td>1.09</td>
<td>1.09</td>
<td>1.10</td>
<td>1.10</td>
</tr>
<tr>
<td>Processor</td>
<td>TPU v2</td>
<td>TPU v3</td>
<td>V100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chip Thermal Design Power (TDP in Watts)</td>
<td>280</td>
<td>450</td>
<td>300</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Measured System Average Power per Accelerator, including memory, network interface, fans, host CPU (W)</td>
<td>208</td>
<td>310</td>
<td>289</td>
<td>288</td>
<td>245</td>
<td>330</td>
</tr>
<tr>
<td>Measured Performance (TFLOPS/s)¹²</td>
<td>24.8</td>
<td>45.6</td>
<td>42.3</td>
<td>48.0</td>
<td>34.4</td>
<td>24.6</td>
</tr>
<tr>
<td>Number of Chips</td>
<td>200</td>
<td>512</td>
<td>1024</td>
<td>1024</td>
<td>1024</td>
<td>10,000</td>
</tr>
<tr>
<td>Training time (days)</td>
<td>6.8</td>
<td>20</td>
<td>30</td>
<td>3.1</td>
<td>27</td>
<td>14.8</td>
</tr>
<tr>
<td>Total Computation (floating point operations)</td>
<td>2.91E+21</td>
<td>4.05E+22</td>
<td>1.12E+23</td>
<td>1.33E+22</td>
<td>8.22E+22</td>
<td>3.14E+23</td>
</tr>
<tr>
<td>Energy Consumption (MWh)</td>
<td>7.5</td>
<td>85.7</td>
<td>232</td>
<td>24.1</td>
<td>179</td>
<td>1,287</td>
</tr>
<tr>
<td>% of Google 2019 total energy consumption (12.2 TWh = 12,200,000 MWh) [Goo20]</td>
<td>0.000006%</td>
<td>0.00070%</td>
<td>0.00190%</td>
<td>0.00020%</td>
<td>0.00147%</td>
<td>0.01055%</td>
</tr>
<tr>
<td>Gross tCO₂e for Model Training</td>
<td>3.2</td>
<td>46.7</td>
<td>96.4</td>
<td>4.8</td>
<td>72.2</td>
<td>552.1</td>
</tr>
<tr>
<td>Net tCO₂e for Model Training</td>
<td>3.2</td>
<td>46.7</td>
<td>96.4</td>
<td>4.3</td>
<td>59.1</td>
<td>552.1</td>
</tr>
</tbody>
</table>
GPT-3 175B model required 3.14E23 FLOPS of computing for training. Even at theoretical 28 TFLOPS for V100 and lowest 3 year reserved cloud pricing we could find, this will take 355 GPU-years and cost $4.6M for a single training run.
President Franklin D. Roosevelt was born in January 1882.

Lily couldn't believe her eyes. The waitress had brought the largest piece of chocolate cake she had ever seen.

Our hand-picked and sun-dried orchard in Georgia.

Peaches are at our orchard in Georgia.

When was Franklin D. Roosevelt born?

1882
Index of /wikidatawiki/entities/

latest-all.json.bz2  27-Jul-2021 11:32  68418560489
latest-all.json.gz   27-Jul-2021 04:58  102963487951

~70GB compressed = 13B float32 parameters
<table>
<thead>
<tr>
<th>Model</th>
<th>NQ</th>
<th>WQ</th>
<th>TQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open-domain SoTA</td>
<td>41.5</td>
<td>42.4</td>
<td>57.9</td>
</tr>
<tr>
<td>T5.1.1-Base</td>
<td>25.7</td>
<td>28.2</td>
<td>24.2</td>
</tr>
<tr>
<td>T5.1.1-Large</td>
<td>27.3</td>
<td>29.5</td>
<td>28.5</td>
</tr>
<tr>
<td>T5.1.1-XL</td>
<td>29.5</td>
<td>32.4</td>
<td>36.0</td>
</tr>
<tr>
<td>T5.1.1-XXL</td>
<td>32.8</td>
<td>35.6</td>
<td>42.9</td>
</tr>
<tr>
<td>T5.1.1-XXL + SSM</td>
<td>35.2</td>
<td>42.8</td>
<td>51.9</td>
</tr>
</tbody>
</table>
Define "middle ear"(x)

Question Answering: Question Query
Barack Obama was born in Hawaii.(x)
Fact Verification: Fact Query

The Divine Comedy (x)
Jeopardy Question Generation: Answer Query

End-to-End Backprop through q and p_θ

Retriever p_η
(Non-Parametric)

Document Index

Generator p_θ
(Parametric)

Marginalize

The middle ear includes the tympanic cavity and the three ossicles. (y)

Question Answering: Answer Generation
supports (y)
Fact Verification: Label Generation
This 14th century work is divided into 3 sections: "Inferno", "Purgatorio" & "Paradiso" (y)

Question Generation

From “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks” by Lewis et al.
From “Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity” by Fedus et al.
The Summer of Language Models 21 (BigScience)

Overview

The "Summer of Language Models 21 🐠" (in short "BigScience") is a one-year long research workshop on very large language models as used and studied in the field of Natural Language Processing and more generally Artificial Intelligence research.

The workshop is

- conducted from May 2021 to May 2022
- with several live sessions/events spread over the year (first session on April 28th, second session planned for end of July). Find all events here.
- with collaborative tasks aimed at creating, sharing and evaluating a very large multilingual dataset and a very large language model as tools for research

From https://bigscience.huggingface.co/
Thanks.

craffel@gmail.com