Building Better Language Models: Insights from BigScience 🌸

Colin Raffel
From https://huggingface.co/spaces/bigscience/SourcingCatalog
8 copies of the model are trained in parallel on a total of 384 GPUs (data parallelism = 8)

DP (data parallelism)

TP (tensor parallelism)

Model parameters are divided across 4 GPUs (tensor parallelism = 4)

The layers of the model are spread across 12 groups of GPUs (pipeline parallelism = 12)

One full copy («replica») of the model takes 48 GPUs

1 GPU - NVIDIA A100 with 80GB of memory

From “BLOOM: A 176B-Parameter Open-Access Multilingual Language Model” by Le Scao et al.
From "Holistic Evaluation of Language Models" by Liang et al.
TriviaQA zero-shot performance

Accuracy

Parameters (Billions)

From “Language Models are Few-Shot Learners” by Brown et al.
<table>
<thead>
<tr>
<th>Title</th>
<th>URL</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paraphrase identification</td>
<td><a href="https://www.usingenglish.com/forum/threads/60200-Do-these-sentences-mean-the-same">https://www.usingenglish.com/forum/threads/60200-Do-these-sentences-mean-the-same</a></td>
<td>Do these sentences mean the same? No other boy in this class is as smart as the boy. No other boy is as smart as the boy in this class.</td>
</tr>
<tr>
<td>Natural Language Inference</td>
<td><a href="https://ell.stackexchange.com/questions/121446/what-does-this-sentence-imply">https://ell.stackexchange.com/questions/121446/what-does-this-sentence-imply</a></td>
<td>If I say: He has worked there for 3 years. does this imply that he is still working at the moment of speaking?</td>
</tr>
<tr>
<td>Summarization</td>
<td><a href="https://blog.nytsoi.net/tag/reddit">https://blog.nytsoi.net/tag/reddit</a></td>
<td>... Lately I've been seeing a pattern regarding videos stolen from other YouTube channels, reuploaded and monetized with ads. These videos are then mass posted on Reddit by bots masquerading as real users. tl;dr: Spambots are posting links to stolen videos on Reddit, copying comments from others to masquerade as legitimate users.</td>
</tr>
<tr>
<td>Pronoun resolution</td>
<td><a href="https://nursecheung.com/ati-teas-guide-to-english-language-usage-understanding-pronouns/">https://nursecheung.com/ati-teas-guide-to-english-language-usage-understanding-pronouns/</a></td>
<td>Jennifer is a vegetarian, so she will order a nonmeat entrée. In this example, the pronoun she is used to refer to Jennifer.</td>
</tr>
</tbody>
</table>
Summarization
The picture appeared on the wall of a Poundland store on Whymark Avenue [...] How would you rephrase that in a few words?

Paraphrase identification
"How is air traffic controlled?" "How do you become an air traffic controller?"
Pick one: these questions are duplicates or not duplicates.

Question answering
I know that the answer to "What team did the Panthers defeat?" is in "The Panthers finished the regular season [...]". Can you tell me what it is?

Multi-task training
Zero-shot generalization

Natural language inference
Suppose "The banker contacted the professors and the athlete". Can we infer that "The banker contacted the professors"?

Graffiti artist Banksy is believed to be behind [...] Not duplicates
Arizona Cardinals
Yes

From “Multitask Prompted Training Enables Zero-Shot Task Generalization” by Sanh et al.
From “Multitask Prompted Training Enables Zero-Shot Task Generalization” by Sanh et al.
From “Multitask Prompted Training Enables Zero-Shot Task Generalization” by Sanh et al.
From “Multitask Prompted Training Enables Zero-Shot Task Generalization” by Sanh et al.
From “Multitask Prompted Training Enables Zero-Shot Task Generalization” by Sanh et al.
From “Crosslingual Generalization through Multitask Finetuning” by Muennighoff et al.
From “Crosslingual Generalization through Multitask Finetuning” by Muennighoff et al.
From “Crosslingual Generalization through Multitask Finetuning” by Muennighoff et al.
From “Crosslingual Generalization through Multitask Finetuning” by Muennighoff et al.
From “Crosslingual Generalization through Multitask Finetuning” by Muennighoff et al.
From “Crosslingual Generalization through Multitask Finetuning” by Muennighoff et al.
Performance on languages that were never intentionally trained on
From “Crosslingual Generalization through Multitask Finetuning” by Muennighoff et al.
Performance on **held-out** tasks

Average zero-shot accuracy on 13 held-out tasks (%)

- Instruction tuning
- Untuned model

$T\theta = 11B$ parameters

From “Fine-Tuned Language Models are Zero-Shot Learners” by Wei et al.
From "What Language Model Architecture and Pretraining Objective Work Best for Zero-Shot Generalization?" by Wang et al.
Causal and non-causal decoder performance is similar when using a non-masked language modeling objective.
Encoder-decoder does very poorly when using a non-masked language modeling objective.
For masked language modeling, the opposite is true - non-causal visibility on the prefix helps a lot!
### Zero-shot
The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1 Translate English to French:  task description
2 cheese =>  prompt
```

### Few-shot
In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1 Translate English to French:  task description
2 sea otter => loutre de mer  examples
3 peppermint => menthe poivrée
4 plush giraffe => girafe peluche
5 cheese =>  prompt
```

### Fine-tuning
The model is trained via repeated gradient updates using a large corpus of example tasks.

```
1 sea otter => loutre de mer  example #1
  ↓
  gradient update

1 peppermint => menthe poivrée  example #2
  ↓
  gradient update
  ↓
  ...
  ↓

1 plush giraffe => girafe peluche  example #N
  ↓
  gradient update

1 cheese =>  prompt
```

From “Language Models are Few-Shot Learners” by Brown et al.
TriviaQA performance

Parameters (Billions)

Accuracy

Few-shot

Zero-shot

From “Language Models are Few-Shot Learners” by Brown et al.
From "Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning", Liu et al. 2022
<table>
<thead>
<tr>
<th>Method</th>
<th>Inference FLOPs</th>
<th>Training FLOPs</th>
<th>Disk space</th>
</tr>
</thead>
<tbody>
<tr>
<td>T-Few</td>
<td>1.1e12</td>
<td>2.7e16</td>
<td>4.2 MB</td>
</tr>
<tr>
<td>T0 [1]</td>
<td>1.1e12</td>
<td>0</td>
<td>0 B</td>
</tr>
<tr>
<td>T5+LM [14]</td>
<td>4.5e13</td>
<td>0</td>
<td>16 kB</td>
</tr>
<tr>
<td>GPT-3 6.7B [4]</td>
<td>5.4e13</td>
<td>0</td>
<td>16 kB</td>
</tr>
<tr>
<td>GPT-3 13B [4]</td>
<td>1.0e14</td>
<td>0</td>
<td>16 kB</td>
</tr>
<tr>
<td>GPT-3 175B [4]</td>
<td>1.4e15</td>
<td>0</td>
<td>16 kB</td>
</tr>
</tbody>
</table>

From "Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning", Liu et al. 2022
Table 2: Top-5 best methods on RAFT as of writing. T-Few is the first method to outperform the human baseline and achieves over 6% higher accuracy than the next-best method.
References

Multitask Prompted Training Enables Zero-Shot Task Generalization
Crosslingual Generalization through Multitask Finetuning
What Language Model Architecture and Pretraining Objective Work Best for Zero-Shot Generalization?
Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning
BLOOM: A 176B-Parameter Open-Access Multilingual Language Model

Please give me feedback:


Thanks!