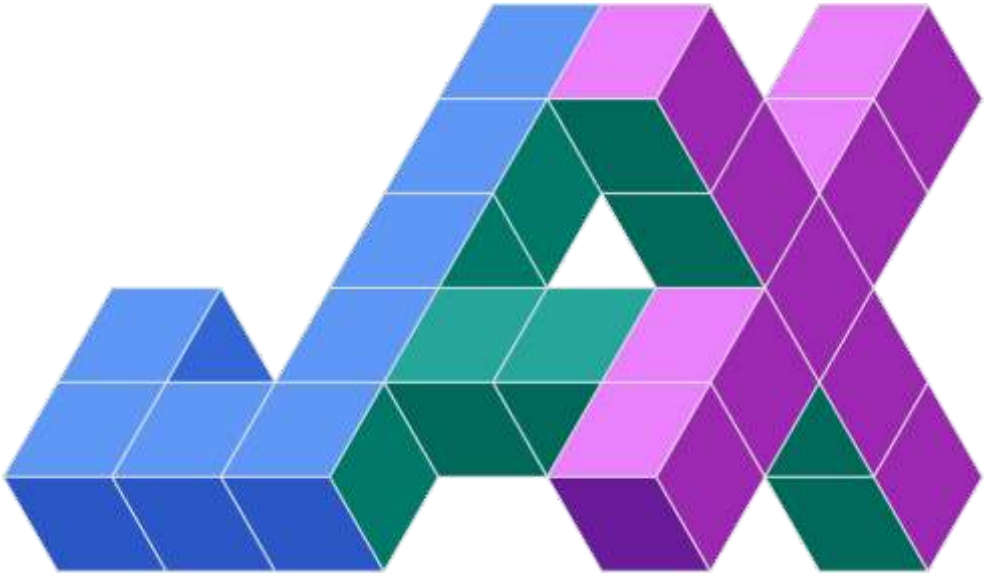
Aran Komatsuzaki

GPT-J-6B: 6B JAX-Based Transformer



Summary:

- We have released GPT-J-6B, 6B JAX-based (Mesh) Transformer LM (Github).
- GPT-J-6B performs nearly on par with 6.7B GPT-3 (or Curie) on various zero-shot down-streaming tasks.
- You can try out this Colab notebook or free web demo.
- This library also serves as an example of model parallelism with xmap on JAX.

Below, we will refer to GPT-J-6B by GPT-J in short.

Why does this project matter?

- GPT-J is the best-performing publicly available Transformer LM in terms of zero-shot performance on various down-streaming tasks.
- GPT-J allows more flexible and faster inference than Tensorflow + TPU counterparts.
- This project required a substantially smaller amount of person-hours than other large-scale model developments did, which demonstrates that JAX + xmap + TPUs is the right set of tools for quick development of large-scale models.

Credit assignment:

- Ben Wang
  - Wrote the code and the Colab notebook, built a part of API and ran experiments.
- Aran Komatsuzaki
  - Proposed this project, designed the high-level plan and the configs, wrote this article and advised Ben.

Acknowledgement:

We would like to thank everyone who have helped this project (in alphabetical order):

- EleutherAI for their general assistance of this project.
- James Bradbury for his valuable suggestions with debugging JAX issues.
- Janko Prester for for creating the web demo frontend.
- Laurence Golding for adding some features to the web demo.
- Leo Gao for running zero shot evaluations for the baseline models for the table.
- TFRC/TRC for providing TPU pods, including TPU v3-256.

Model design:

Our model design and hyperparameter choice closely follow those of 6.7B GPT-3 with some differences. Notably,

- The model was trained on 400B tokens from The Pile dataset with 800GB text.
- Efficient attention (linear, local/sliding window, etc) was not used for simplicity, as it would not have significantly improved throughput at this scale.
- The dimension of each attention head is set to 256, which is twice larger than that of GPT-3 of comparable size. This noticeably improved the throughput with minimal performance degradation.

We have made two minor architectural improvements:

- Rotary embedding for slightly better performance.
- Placing the attention layer and the feedforward layer in parallel for decreased communication.

Performance:

Model	Training FLOPs	LAMBADA PPL ↓	LAMBADA Acc ↑	Winogrande ↑	Hellaswag ↑	PIQA ↑
GPT-2-1.5B	-----	10.63	51.21%	59.4%	50.9%	70.8%
GPTNeo-2.7B	6.8e21	5.63	62.2%	56.5%	55.8%	73.0%
GPT-3-1.3B	2.4e21	5.44	63.6%	58.7%	54.7%	75.1%
GPT-3-Babbage	-----	5.58	62.4%	59.0%	54.5%	75.5%
GPT-3-2.7B	4.8e21	4.60	67.1%	62.3%	62.8%	75.6%
GPT-J-6B	***1.5e22***	***3.99***	***69.7%***	***65.3%***	***66.1%***	***76.5%***
GPT-3-6.7B	1.2e22	4.00	70.3%	64.5%	67.4%	78.0%
GPT-3-Curie	-----	4.00	69.3%	65.6%	68.5%	77.9%
GPT-3-175B	3.1e23	3.00	76.2%	70.2%	78.9%	81.0%
GPT-3-Davinci	-----	3.0	75%	72%	78%	80%

- The zero-shot performance is roughly on par with GPT-3 of comparable size, and the performance gap from GPT-3 of comparable size is closer than the GPT-Neo models.

- Throughput of the 6B GPT-J for training (151k tokens/s) is faster than the 2.7B GPT-Neo (148k tokens/s) on the same hardware (TPU v3-256 pod), demonstrating an approximately 125% improvement in efficiency.
- At the 6B config on a TPU V3-256 pod, GPT-J achieves high absolute efficiency. The hardware has a theoretical maximum of 13.4PFLOPs, and GPT-J achieves 5.4 PFLOPs as measured in the GPT3 paper (ignoring attention computation, ignoring compute-memory tradeoffs like gradient checkpointing). When taking these additional factors into account, 8.1 PFLOPs, or approximately 60% of the theoretical maximum is utilized.
- GPT-J training took roughly five weeks with TPU v3-256.

Sample Outputs:

Below, we will list some interesting samples we found that were generated with GPT-J along with their prompt.

Arithmetic

```
 Prompt -------

What is 75*10?

--------------

Output:

750

What is -0.002 take away 72.75?

-72.752

Calculate -0.5 - 1039.

-1039.5

What is the difference between -1360 and 2?

1362

What is -27.95 less than -20?

7.95

Calculate -0.3 + -169.

-169.3

What is 0.7 minus 0.05?

0.65

Calculate -2 + 0.0899.

-1.9101
```

The model performs addition and subtraction perfectly in this setting of creating the questions by themselves.

Theorem Proving

```
Prompt -------

Theorem 3.2: Any cyclic group is abelian.

Proof:

--------------

Output:

Let G be a cyclic group. If the order is 1, then it is trivial.
If the order is 1, then G has only one element, so it is abelian.

If the order is not 1, let G be a cyclic group of order n.

Then G is generated by a single element a. Let a = gt.

Note that G is an abelian group, so gt^n = e. But t^n = e, so
gt^n = t^n = e, so t^n = g^n.

For any element g in G, (gt)^n = t^n, so g^n = (gt)^n = t^n =
g^n.

So G is abelian.
```

While the model can imitate the style of proving this simple theorem to some extent, there is still a huge gap with human-level accuracy.

NLU

```
Prompt ------

Passage: Badgers are short-legged omnivores in the family
Mustelidae, which also includes the otters, polecats, weasels,
and wolverines. They belong to the caniform suborder of
carnivoran mammals. The 11 species of badgers are grouped in
three subfamilies: Melinae (Eurasian badgers), Mellivorinae (the
honey badger or ratel), and Taxideinae (the American badger). The
Asiatic stink badgers of the genus Mydaus were formerly included
within Melinae (and thus Mustelidae), but recent genetic evidence
indicates these are actually members of the skunk family, placing
them in the taxonomic family Mephitidae.

Question: is a wolverine the same as a badger?

--------------

Output (greedy):

A: The wolverine is a member of the Mustelidae family, but it is
not a badger.

--------------

Output (p = 0.9):

A:

You'll first need to establish what you mean by a wolverine and
what you mean by a badger. Wolverines are fur-bearing members of
the canid order Carnivora, which includes dogs, foxes, and
weasels. They are not the same thing as badgers, and neither are
badgers what wolverines were originally described as. They're
different animals. Badgers are canids; wolverines are canids.
As to what the difference is, there is none.
```

Completion on a question from BoolQ (SuperGLUE). While both sampling methods result in the same correct conclusion, the nucleus sampling hallucinates and contains incorrect reasoning, while the greedy sampling answers concisely and reasonably. In general, we observed that greedy sampling is more accurate and contains less hallcinations than nucleus sampling when the output is supposed to be short like this, which is predictable given that classification task is usually done with greedy sampling.

Coding

```
 Prompt ------

class MultiHeadedAttention(nn.Module):
    def __init__(self, h, d_model, dropout=0.1):

-------------

Output:

super(MultiHeadedAttention, self).__init__()
        self.dropout = nn.Dropout(dropout)
        self.h = nn.ModuleList([nn.Linear(d_model, h) for _ in
range(h)])
        self.d = nn.Linear(d_model, d_model)
        self.d_model = d_model

    def forward(self, x, h_att_mask, return_h_att_mask=False):
        """
        :param x: a sequence of word tokens of shape (batch, seq,
vocab),
        :param h_att_mask: a mask of attention heads, with shape
(batch, seq, head)
        :param return_h_att_mask: whether to return the attention
mask
        """
```

While the model knows what attention is like, it does not know its precise mechanism as in theorem proving.

Geese

```
 Prompt ------

Fun Facts About Geese:
1. Geese have impressive visual capabilities!

-------------

Output:

They can see a human face and react to it.
2. Geese produce a lot of noise!
They can be heard and seen flying over 100 miles away!
3. Geese are the biggest birds of prey in the world!
They have an average wingspan of 45 inches.
4. Geese can fly 1,000 miles per hour!
They can travel up to 40 miles per hour for short distances.
5. Geese usually mate for life!
They will mate with any other goose with which they are
closely associated.
6. Geese have a variety of vocalizations.
The most prominent one is the honking call, which is
used mainly for communication.
7. Geese can be found on every continent except Antarctica.

Geese:
Geese are birds of the genus Anser. They are native to the
Northern Hemisphere, and
are found in a wide variety of environments.
Geese are typically very solitary birds and seldom form breeding
pairs.
Geese are characterized by their long necks, distinctive head
crests, and heavy body.
In temperate regions, geese are a common sight, for they form
large flocks.
The most common of these
```

Links:

- Github repository for GPT-J
- Colab notebook
- Web demo

Citation:

To cite GPT-J-6B pretrained model:

```
@misc{gpt-j,
  author = {Wang, Ben and Komatsuzaki, Aran},
  title = {{GPT-J-6B: A 6 Billion Parameter Autoregressive
Language Model}},
  howpublished = {\url{https://github.com/kingoflolz/mesh-
transformer-jax}},
  year = 2021,
  month = May
}
```

To cite Mesh Transformer JAX, the codebase used for the training:

```
@misc{mesh-transformer-jax,
  author = {Wang, Ben},
  title = {{Mesh-Transformer-JAX: Model-Parallel Implementation
of Transformer Language Model with JAX}},
  howpublished = {\url{https://github.com/kingoflolz/mesh-
transformer-jax}},
  year = 2021,
  month = May
}
```


Published by Aran Komatsuzaki

View all posts by Aran Komatsuzaki

June 4, 2021
Uncategorized

COMMENTS ARE CLOSED.

UP ↑