Sheared-LLaMA: Accelerating Language Model Pre-Training Via Structured Pruning
Mengzhou Xia1, Tianyu Gao1, Zhiyuan Zeng2, Danqi Chen1
1Princeton University, 2Tsinghua University
mengzhou@princeton.edu @xiamengzhou

Open-source development of LLMs

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Size</th>
<th>When?</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLaMA-1</td>
<td>7B, 13B, 34B, 70B</td>
<td>02/2023</td>
<td>1T/1.4T tokens</td>
</tr>
<tr>
<td>INCITE-Base</td>
<td>3B, 7B</td>
<td>05/2022</td>
<td>800M tokens on RedPajama</td>
</tr>
<tr>
<td>OpenLLaMA-v1</td>
<td>7B, 13B</td>
<td>06/2023</td>
<td>1T tokens on RedPajama</td>
</tr>
<tr>
<td>OpenLLaMA-v2</td>
<td>7B, 13B</td>
<td>07/2023</td>
<td>1T RP, RefinedWeb, StarCoder</td>
</tr>
<tr>
<td>LLaMA-2</td>
<td>7B, 13B, 70B</td>
<td>07/2023</td>
<td>2T Unknown tokens</td>
</tr>
<tr>
<td>Mistral</td>
<td>7B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vi</td>
<td>6B, 34B</td>
<td>11/2023</td>
<td>Unknown</td>
</tr>
</tbody>
</table>

• Fast development of open-source intermediate-sized models
• All (could be) trained from scratch! Expensive :(  

Prune a large model to a small model: LLM-Shearing Method

• How to prune? Targeted structured pruning prunes models to a pre-specified shape
  
  - Placing masks for each head, intermediate dimension, hidden dimension, layer
  - The masks are learned with L0 regularization with Lagrange constraints

• Prune from LLaMA2-7B to 1.3B and 2.7B
  
  - Prune with 0.4B tokens, continue pre-training with 50B tokens
  - Outperforms existing pretraining models on 14 downstream tasks (zero-shot)

Pruning

- When building a series of models, start training the large one and prune it down.
- When building small-scale domain specific models, prune the keep pre-training.

Effect of Dynamic Batch Loading

- The loss disparity is much more even across domains
- Downstream is better with dynamic batch loading

Generality

- Prune a large model to a small model: LLM-Shearing Method

Results

- Prune with 0.4B tokens, continue pre-training with 50B tokens
- Outperforms existing pretraining models on 14 downstream tasks (zero-shot)

You should never train 1.3B and 2.7B from scratch ever again! Pruning is always more efficient!

Outperforms other models on instruction tuning — pruning does not destroy model's generation ability

In our paper, please find:

- Targeted structured pruning delivers submodels that are fast at inference and better in performance compared to LLM-Pruner (Ma et al., 2023)

Research Question

Can we produce stronger LLMs more efficiently given the existing strong LLMs?

- YES, WE CAN! Via Structured Pruning!

- Pruning LLaMA2-7B to 2.7B w/ 50B tokens (32x less tokens)
- Outperforms strong existing models
- No sign of convergence, could be further improved

Pruning

- What data to use? How about simply using a pre-training dataset such as RedPajama?
- Domain reference loss: a hypothetical target model loss on a held-out validation set estimated by scaling law
- Presents a loss disparity between the pruned model and the target loss
- Dynamic batch loading: load data dynamically across domains to reach the target loss simultaneously
- Update rule: For every K batch, update the loading proportion with the following rule
  
  \[
  \alpha_t = w_{t=m} \cdot \exp(\Delta_t); \quad w_t = \frac{\alpha_t}{\sum_i \alpha_i(t)}
  \]

- Does not require training a reference model or a proxy model as Xie et al., 2023

Generality

- Train a model to match the performance of any other model.
- Retain performance of some domains when continue training on a new domain.