Efficient Training of BERT by Progressively Stacking

Linyuan Gong, Di He, Zhuohan Li, Tao Qin, Liwei Wang, Tie-Yan Liu

Peking University & Microsoft Research Asia

ICML | 2019
### BERT: Effective Model with Huge Costs

<table>
<thead>
<tr>
<th>Model</th>
<th>Data</th>
<th>Training</th>
</tr>
</thead>
<tbody>
<tr>
<td>110M/330M params</td>
<td>3.4B words (enwiki + book)</td>
<td>128K tokens * 1M updates</td>
</tr>
</tbody>
</table>

4 Days on 4 TPUs or 23 Days on 4 Tesla P40 GPUs
Attention Distributions of BERT

Neighborhood & [CLS]

Similar!

High-level layers

L8 H1
L10 H12
L12 H10

Low-level layers

L2 H5
L4 H5
L6 H12
Stacking

Efficient Training of BERT by Progressively Stacking
Stacking Progressively

\[ M_0' \leftarrow \text{InitBERT}(L/2^k) \]
\[ M_0 \leftarrow \text{Train}(M_0') \quad \{ \text{Train from scratch.} \} \]
\[ \text{for } i \leftarrow 1 \text{ to } k \text{ do} \]
\[ M_i' \leftarrow \text{Stack}(M_i) \quad \{ \text{Doubles the number of layers.} \} \]
\[ M_i \leftarrow \text{Train}(M_i') \quad \{ M_i \text{ has } L/2^{k-i} \text{ layers.} \} \]
\[ \text{end for} \]
\[ \text{return } M_k \]

6/12/2019
Efficient Training of BERT by Progressively Stacking
Result

![Graph showing valid loss over wall time (hours). The graph compares stacking, identity+noise, identity, and baseline methods. The baseline method shows a significant improvement in valid loss with wall time, indicating an approximate 25% reduction.](image-url)
Result
## Result

<table>
<thead>
<tr>
<th></th>
<th>CoLA</th>
<th>SST-2</th>
<th>MRPC</th>
<th>STS-B</th>
<th>QQP</th>
<th>MNLI</th>
<th>QNLI</th>
<th>RTE</th>
<th>GLUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT-Base</td>
<td>52.1</td>
<td>93.5</td>
<td>88.9/84.8</td>
<td>87.1/85.8</td>
<td>71.2/89.2</td>
<td>84.6/83.4</td>
<td>90.5</td>
<td>66.4</td>
<td>78.3</td>
</tr>
<tr>
<td>Stacking</td>
<td><strong>56.2</strong></td>
<td><strong>93.9</strong></td>
<td><strong>88.2/83.9</strong></td>
<td><strong>84.2/82.5</strong></td>
<td><strong>70.4/88.7</strong></td>
<td><strong>84.4/84.2</strong></td>
<td><strong>90.1</strong></td>
<td><strong>67.0</strong></td>
<td><strong>78.4</strong></td>
</tr>
</tbody>
</table>
Take aways

• Progressively stacking training for BERT is efficient
  • [https://github.com/gonglinyuan/StackingBERT](https://github.com/gonglinyuan/StackingBERT)
  • Poster #50

• Towards a better understanding of Transformer
  • Codes and model ckpts @ [https://github.com/zhuohan123/macaron-net](https://github.com/zhuohan123/macaron-net)