Overview

Problem

State-of-the-art machine translation methods employ massive amounts of parameters. Compressing such models is essential for efficient inference on edge-devices.

Proposition

We propose a quantization strategy for the Transformer. Our goal is to:

▶ Quantize all operations which can provide a computational speed gain.
▶ Exploit hardware resources as efficiently as possible.
▶ Maintain accuracy with respect to full-precision.

FullyQT

![Fully Quantized Transformer](Image)

**Quantization** [4]

Given an element $x$ of a tensor $X$, we apply the quantization function $Q$:

$$Q(x) = \left\lfloor \frac{\text{clamp}(x, x_{\text{min}}, x_{\text{max}}) - x_{\text{min}}}{s} \right\rfloor s + x_{\text{min}}$$

(1)

For weights:

$$x_{\text{min}} = \min(X) \quad x_{\text{max}} = \max(X)$$

For activations:

$$x_{\text{min}}, x_{\text{max}} = \text{Running Estimates}$$

Pruning

![Pruning](Image)

Table 4: Adaptive vs fixed rate pruning, equal proportions.

<table>
<thead>
<tr>
<th>Model</th>
<th>Method</th>
<th>BLEU</th>
<th>EN-DE</th>
<th>EN-FR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>No pruning</td>
<td>27.09</td>
<td>55.64</td>
<td>28.74</td>
</tr>
<tr>
<td>No bucketing</td>
<td>27.09</td>
<td>55.64</td>
<td>28.74</td>
<td></td>
</tr>
<tr>
<td>No gradient clipping</td>
<td>27.09</td>
<td>55.64</td>
<td>28.74</td>
<td></td>
</tr>
<tr>
<td>No LayerNorm Denominator Quantization</td>
<td>27.09</td>
<td>55.64</td>
<td>28.74</td>
<td></td>
</tr>
<tr>
<td>8-bit</td>
<td>Quantized Weights, Full-precision Activations</td>
<td>27.09</td>
<td>55.64</td>
<td>28.74</td>
</tr>
</tbody>
</table>

References


