motivation
- The best pre-trained NLP models are extremely large
- This limits research to those who can afford large GPUs
- Model compression of Transformers offers a solution
- Only pruning / knowledge distillation has been tried, and very recently
- Quantization and compressed self-attention analysis have not been conducted


dataset
- WMT English-German Translation
- 4.5 million sentences to train on and approximately 3000 sentences to validate and test on
- BLEU score to evaluate performance/compression trade-off

flexible binary scheme (BS-Flex)
- Improve on a 1-bit quantization scheme [2]
- Allows reassignment of clusters during retraining unlike K-Means
- Set values of weights based on the average of the two centroids $c_1$ and $c_2$

\[
\begin{align*}
\text{assign}_{i,j} = \begin{cases} 
  c_1 & \text{if } w_{i,j} > \frac{c_1 + c_2}{2} \\
  c_2 & \text{if } w_{i,j} \leq \frac{c_1 + c_2}{2} 
\end{cases}
\end{align*}
\]
- We also experimented with a scheme that fixed the centroids, which is the original way [2].

selected results

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU</th>
<th>% Perf</th>
<th>CR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Transformer</td>
<td>28.09</td>
<td>100</td>
<td>1x</td>
</tr>
<tr>
<td>K-means (KM) 4-bit</td>
<td>27.65</td>
<td>98.43</td>
<td>5.85x</td>
</tr>
<tr>
<td>KM 1-bit (self-att only)</td>
<td>12.07</td>
<td>42.96</td>
<td>23.37x</td>
</tr>
<tr>
<td>KM 1-bit (self-att only)</td>
<td>24.96</td>
<td>88.85</td>
<td>10.02x</td>
</tr>
<tr>
<td>BS-Flex (self-att only)</td>
<td>25.54</td>
<td>90.92</td>
<td>10.02x</td>
</tr>
<tr>
<td>Pruning 30-&gt;50%</td>
<td>26.40</td>
<td>93.98</td>
<td>2x</td>
</tr>
<tr>
<td>Pruning 50-&gt;80%</td>
<td>25.02</td>
<td>89.07</td>
<td>5x</td>
</tr>
</tbody>
</table>

- Compression works really well!
- Our binarization scheme performs as well or better than K-Means and is more than 10x faster due to PyTorch indexing
- Pruning seems to do much worse than quantization, harder to train
- Binarizing only attention layers still gives 90% of performance!

effects on self-attention?

From top to bottom, the rows represent original, 8-bit, 4-bit, and BS-Flex (att only)

conclusion
- Significant Transformer compression can be achieved with minimal loss in performance
- This demonstrates the potential for quality compression of state-of-the-art NLP architectures
- Our binarization scheme runs faster and performs better than standard 1-bit quantization
- Self-attention is highly resistant to quantization; replacing elements with only 16 values produces a nearly identical distribution

references
[1] Song Han, Huizi Mao, and William J. Dally. “Deep Compression: Compressing Deep Neural Network with Pruning, Trained Quantization and Huffman Coding”.