Overview

- **Multi-head** attention is central to modern state-of-the-art NLP models.
- But to what extent do we actually need multiple heads?
- Key findings:
  - Not all heads are needed at test time.
  - A significant amount of heads can be pruned, and this phenomenon appears early in training.
  - Multi-headedness is more important in some parts of the model than others.

Multi-Head Attention

As a reminder, the attention mechanism parametrized by $W_k, W_q, W_v, W_o \in \mathbb{R}^{d \times d}$ computes the weighted sum:

$$\text{Att}_{k,q,v,w}(x, q) = W_o \sum_{i=1}^{n} \alpha_i W_k x_i$$

where $\alpha_i = \text{softmax} \left( \frac{q^T W_q^T W_k x_i}{\sqrt{d}} \right)$

In multi-head attention, $N_h$ independently parameterized attention layers are applied in parallel:

$$\text{MHAtt}(x, q) = \sum_{h=1}^{N_h} \text{Att}_{k,q,v,w}(x, q)$$

We can even go beyond that and remove all but one head from each individual layer:

### Systematic Pruning

We prune head $h$ according to the importance score given by the expected sensitivity of the loss function to the value of the mask variable $\xi_h$

$$I_h := \mathbb{E}_{x \sim X} \left\{ \frac{\partial L(x)}{\partial \xi_h} \right\} = \mathbb{E}_{x \sim X} \text{Att}(x)^T \frac{\partial L(x)}{\partial \text{Att}(x)}$$

### Removing Attention Heads

Most heads can be individually removed without significantly impacting performance.

**Figure:** Distribution of heads by model score after masking.

<table>
<thead>
<tr>
<th>Layer</th>
<th>Enc-Enc</th>
<th>Enc-Dec</th>
<th>Dec-Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1.31</td>
<td>0.24</td>
<td>-0.03</td>
</tr>
<tr>
<td>2</td>
<td>-0.16</td>
<td>0.06</td>
<td>0.12</td>
</tr>
<tr>
<td>3</td>
<td>0.12</td>
<td>0.05</td>
<td>0.18</td>
</tr>
<tr>
<td>4</td>
<td>-0.15</td>
<td>-0.24</td>
<td>0.17</td>
</tr>
<tr>
<td>5</td>
<td>0.02</td>
<td>-1.55</td>
<td>-0.04</td>
</tr>
<tr>
<td>6</td>
<td>-0.36</td>
<td>-13.56</td>
<td>0.24</td>
</tr>
</tbody>
</table>

**Table:** Best delta BLEU by layer when only one head is kept in the WMT model. Underlined ⇔ statistically significant ($p < 0.01$).

**Figure:** Evolution of score by number of heads pruned by $I_h$ (blue) and oracle score difference (dashed green).

**Figure:** BERT when incrementally pruning heads from each attention type in the WMT model.

**Figure:** Relationship between percentage of heads pruned and relative score decrease during training of a transformer on IWSLT De-En.

---

Code: https://github.com/pmichel31415/are-16-heads-really-better-than-1