On the Expressivity Role of LayerNorm in Transformers’ Attention
Shaked Brody, Uri Alon, Eran Yahav

Code: https://github.com/tech-srl/layer_norm_expressivity_role

Known so far
LayerNorm is the least studied component of Transformers
- Normalizing activations in the forward pass
- Normalizing gradients in the backward pass

This work: A Geometric Interpretation of LayerNorm
LayerNorm can be seen as a **projection** followed by a **scaling** operation

Projection
Helps to compute even simple tasks by learning to attend equally to all keys
- Helps to compute simple tasks by learning to attend equally to all keys.
- "Majority" task: Predict the most frequent token

Scaling
- The fraction of "unselectable" keys in different layers of a language model
- Interior key vectors cannot receive the highest attention score
- Convex Hull

<table>
<thead>
<tr>
<th>Model</th>
<th>$L_1$</th>
<th>$L_2$</th>
<th>$L_3$</th>
<th>$L_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o scaling</td>
<td>51.0</td>
<td>32.2</td>
<td>34.7</td>
<td>36.8</td>
</tr>
<tr>
<td>w/ scaling</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>