On the Expressivity Role of LayerNorm in Transformers’ Attention

Shaked Brody
Technion

Uri Alon
Language Technologies Institute
Carnegie Mellon University

Eran Yahav
Technion
LayerNorm - Known so far

LayerNorm is the least studied component of Transformers

Normalizing activations in the **forward** pass

Normalizing gradients in the **backward** pass

Attention Is All You need, Vaswani et al., 2017
This work: A Geometric Interpretation of LayerNorm

projection

scaling
LayerNorm projects the key vectors onto the same hyperplane
LayerNorm – Projection

“Majority” task: Predict the most frequent token

Angle to $\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$

Training Loss

$\angle$ \hspace{1cm} $\angle$
LayerNorm – Scaling

“Unselectable” keys problem [Demter et al., 2020; Grivas et al., 2022]

Interior key vectors cannot receive the highest attention score
LayerNorm – Scaling

The fraction of “unselectable” $n$ vectors of dimension $d$
LayerNorm – Scaling

The fraction of “unselectable” key vectors in different layers of a language model

<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>
Summary

- LayerNorm is crucial for the expressivity of attention in Transformers
- LayerNorm can be seen as a projection followed by a scaling operation
- Projection helps the model to learn to attend equally to all keys
- Scaling eliminates the “unselectable” keys problem

shakedbr@cs.technion.ac.il

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