GOAL: An attempt to understand the DNNs from a statistical perspective

HOW: Statistical properties of bottleneck (BN) layer pre-activations (Z) and activations (Y) are studied.

CONTRIBUTIONS:
1. Distribution of the NN activation in the BN layer was analytically derived.
2. Statistical properties of the BN features were empirically studied and compared with analytic pdf.
3. Sparsity of ReLU was (re-)explained.
4. Post-processing of the BN features through statistical normalisation for ASR were investigated.

EXPERIMENTS: Aurora-4, train by clean/additive.

RESULTS: Up to 2% absolute (9% relative) performance gain (WER reduction) was achieved in mismatch condition.

STATISTICAL DISTRIBUTION OF BOTTLENECK FEATURES

Statistical properties of bottleneck (BN) layer pre-activations (Z) and activations (Y) are studied.

Assumptions for Approximating $P_z(z)$

1. Central Limit Theorem (CLT)
   \[ z \sim \mathcal{N}(z; \mu_z, \sigma_z^2) \]
2. \( \text{Prob}(z > 0) \approx \text{Prob}(z < 0) \)
   \[ \mu_z \rightarrow 0 \quad z \sim \mathcal{N}(z; 0, \sigma_z^2) \]

Density Estimating for Nodes with tanh Activation

\[ P_{Y_{tanh}}(y) = \frac{1}{1 - y^2} \phi \left( \frac{1 + y}{\sigma_z} - \frac{1}{2\sigma_z} \log \frac{1 + y}{1 - y} \right) \]

Density of Sigmoid

\[ P_{Y_{sigmoid}}(y) = \frac{1}{\sigma_z} \phi \left( \frac{1 - y}{\sigma_z} \right) \]

Non-linearity of NNs and Density Shape Parameter

\[ \sigma_z < 1 \rightarrow \text{Nodes/NN operates in linear mode} \]
\[ \sigma_z > 1 \rightarrow \text{Nodes/NN operates in non-linear mode} \]

EMPIRICAL STUDIES

1. Zero mean approximation for Z is reasonable.
2. \( \sigma_z > 1 \) \( \Rightarrow \) DNN operates in the non-linear mode.
3. Distribution of Z can be easily fitted by a GMM.
4. Distribution of Y may NOT be fitted by a GMM.
5. DNN decorrelates the features in the BN layer.
6. Distribution of Y matches the derived equation.

STATISTICAL NORMALISATION OF THE BOTTLENECK (BN) FEATURES

Experimental Results

Table 1: WER for Aurora-4 (Kaldi-LDA-MLLT).

<table>
<thead>
<tr>
<th>Feature</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>Ave4</th>
</tr>
</thead>
<tbody>
<tr>
<td>BN (baseline)</td>
<td>3.87</td>
<td>7.96</td>
<td>21.80</td>
<td>32.72</td>
<td>16.58</td>
</tr>
<tr>
<td>BN+MN</td>
<td>3.64</td>
<td>7.66</td>
<td>21.02</td>
<td>32.20</td>
<td>16.13</td>
</tr>
<tr>
<td>BN+MN</td>
<td>4.07</td>
<td>8.31</td>
<td>20.34</td>
<td>33.04</td>
<td>16.44</td>
</tr>
<tr>
<td>BN+Gauss</td>
<td>4.15</td>
<td>8.12</td>
<td>20.18</td>
<td>32.67</td>
<td>16.28</td>
</tr>
<tr>
<td>BN+HEQ</td>
<td>3.96</td>
<td>7.43</td>
<td>19.76</td>
<td>30.87</td>
<td>15.50</td>
</tr>
<tr>
<td>BN+PCA</td>
<td>3.75</td>
<td>7.88</td>
<td>21.56</td>
<td>32.46</td>
<td>16.41</td>
</tr>
<tr>
<td>BN+DCT</td>
<td>3.77</td>
<td>7.77</td>
<td>21.76</td>
<td>32.49</td>
<td>16.44</td>
</tr>
</tbody>
</table>