Training and Inference on Any-Order Autoregressive Models the Right Way

Andy Shih, Dorsa Sadigh, Stefano Ermon
andyshih@cs.stanford.edu

Joint Likelihood

Joint Likelihood

\[ \log p(x) = \sum_{i=1}^{N} \log p(x_{i} | x_{<i}) \]

Any-Order Autoregressive Models

Any-Order Autoregressive Models learn all orderings

\[ p(x_{3} | x_{2}, x_{4}) \]
\[ p(x_{3} | x_{1}) \]
\[ p(x_{2} | x_{3}) \]

All univariate conditionals (edges) learned with a single weight-tied neural network

Input x ─── model ─── p(x_i | x_e)

MAC: Mask-Tuned Arbitrary Conditional Models

Two improvements:

1. For each node, we only need one path from ∅
2. Only need the red edges!

Some edges will be evaluated more frequently!

Text8 dataset (bpd, lower is better)

<table>
<thead>
<tr>
<th>Model</th>
<th>joint</th>
<th>marginal</th>
</tr>
</thead>
<tbody>
<tr>
<td>O3-Transformer</td>
<td>1.64</td>
<td>1.47</td>
</tr>
<tr>
<td>D3PM</td>
<td></td>
<td>1.43</td>
</tr>
<tr>
<td>ARDM (14000 epochs)</td>
<td>1.48</td>
<td>1.12</td>
</tr>
<tr>
<td>ARDM (3000 epochs)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAC (3000 epochs)</td>
<td>1.40</td>
<td>1.09</td>
</tr>
</tbody>
</table>