1. Motivation

- Training **deep networks** has been widely adopted and has **shown effectiveness** in image recognition, QA and text classification.
- Very deep and effective model training still **remains challenging for NMT**.

2. Framework

3. Depth Growing

4. Two-stage Training

5. Experiments

Overall Results

- WMT14 En→De and WMT14 En→Fr

<table>
<thead>
<tr>
<th>Model</th>
<th>En→De</th>
<th>En→Fr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transformer (6B)†</td>
<td>28.40</td>
<td>41.80</td>
</tr>
<tr>
<td>Transformer (6B)</td>
<td>28.91</td>
<td>42.69</td>
</tr>
<tr>
<td>Transformer (8B)</td>
<td>28.75</td>
<td>42.63</td>
</tr>
<tr>
<td>Transformer (10B)</td>
<td>28.63</td>
<td>42.73</td>
</tr>
<tr>
<td>Transparent Attn (16B)†</td>
<td>28.04</td>
<td>–</td>
</tr>
<tr>
<td>Ours (8B)</td>
<td><strong>29.92</strong></td>
<td><strong>43.27</strong></td>
</tr>
</tbody>
</table>

*dagger: results reported in previous works*

- We achieve **30.07 BLEU score** on En→De with 10 blocks (10B).

Analysis

- **Directly Stacking (DS):** extend the 6-block baseline to 8-block by directly stacking 2 blocks.
- **Ensemble Learning (Ensemble):** separately train 2 models and ensemble their decoding results.

The test performances of WMT14 En→De translation task.

**Code**
- [https://github.com/apeterswu/Depth_Growing_NMT](https://github.com/apeterswu/Depth_Growing_NMT)

**Contact**
- wulijun3@mail2.sysu.edu.cn (SYSU)
- yingce.xia@microsoft.com (MSRA)