Beyond Transformer

- **Self-attentive feed-forward** sequence models achieve impressive results on sequence modeling tasks, making them a compelling alternative to RNNs.
- The **Transformer** relies entirely on a self-attention mechanism to compute a series of context-informed vector-space representations of the symbols in its input and output.
  - Straightforward to parallelize
  - Global receptive field (compared to convolutional architectures)
- Transformer foregoes the RNN’s **inductive bias** towards learning iterative or recursive transformations.
  - Appears to be crucial for several algorithmic and language understanding tasks
- Due to fixed depth, the Transformer is not **computationally universal**.

Universal Transformer with Dynamic Halting

- The number of parallel steps need not be fixed:
  - Adaptive Computation Time (Graves, 2016)
  - Allows applying different transformations to different parts of the input:
    - Can allocate more processing steps to symbols that require more computations, e.g. ambiguous words.
    - Allows applying different transformations to different inputs:
      - E.g. more steps for longer sequences.

Universal Transformers

- **Parallel-in-Time Recurrence**
  - Iteratively refines its representations for all positions in the sequence in parallel
  - Retains parallelizability and global receptive field of the Transformer model
  - But with recurrent inductive bias of RNNs
  - Computationally universal (under certain assumptions)
    - Can reduce Neural GPU to UT

 Experiments

- **bAbI Question Answering Tasks**

<table>
<thead>
<tr>
<th>Model</th>
<th>Easy</th>
<th>Difficult</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original RNN</td>
<td>0.95</td>
<td>0.87</td>
</tr>
<tr>
<td>LSTM</td>
<td>0.93</td>
<td>0.90</td>
</tr>
<tr>
<td>Universal Transformer</td>
<td>0.95</td>
<td>0.87</td>
</tr>
</tbody>
</table>

- **Algorithmic/ Learning to Execute Tasks**

- **Subject-Verb Agreement Task**

- **LAMBADA Language Modeling**