Motivation

Recurrent networks scale poorly

- The computation of $c_t$ is suspended until $c_{t+1}$ becomes completely available.
- This sequential dependency breaks computation into a successive execution of relative small computation for each $c_t$.
- As a result, RNNs cannot utilize the full parallelization power of hardware and runs much slower than attention and convolution.

Contribution

Simple Recurrent Unit (SRU), a recurrent unit that is no longer a parallelization bottleneck.

- exhibits the same parallelism as convolution and attention.
- retains modeling capacity as LSTM and GRU etc.

Open source code: https://github.com/OpenNMT/sru

SRU

Basic architecture

SRU involves relatively few computation, which decomposes into two sub-components:

(i) light gated recurrence

$$f_t = \sigma(W_x x_t + v_t \odot c_{t-1} + b_f)$$
$$c_t = f_t \odot c_{t-1} + (1 - f_t) \odot (W_h x_t)$$

(ii) highway connection

$$r_t = \sigma(W_x x_t + v_t \odot c_{t-1} + b_r)$$
$$h_t = r_t \odot c_{t-1} + (1 - r_t) \odot x_t$$

We use element-wise multiplication (e.g., $v_f \odot c_{t-1}$) for hidden-to-hidden connection.

Optimizations

The architecture enables two optimizations that achieve significant speed-up over traditional RNNs: (i) group matrix multiplications across all steps into one single multiplication, and (ii) write a custom function to perform the element-wise operations for computation intensity.

SRU vs. LSTM

While LSTM also uses a light gated recurrence from $g_{t-1}$ to $c_t$, it uses a full recurrence from $c_t$ to $g_{t+1}$ which intuitively seems wasteful.

$$g \in \{f, r, i, o\}$$ is a gate, full: $g_t = \sigma(W_s x_t + V_g c_{t-1} + b_g)$

light: $g_t = \sigma(W_s x_t + V_0 c_{t-1} + b_g)$

Standard NN uses matrix multiplications to stack layers. SRU uses highway connections shown to be effective in ResNet/highway networks.

Results

Ablation analysis

Successively disable components in SRU to confirm the impact of our design choices.

Classification

Tested on 6 sentence classification benchmarks. SRU operates 5-9x faster than cuDNN LSTM, achieving on par or better results than various baselines.

<table>
<thead>
<tr>
<th>Model</th>
<th>4 layers</th>
<th>6 layers</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRU (full)</td>
<td>70.7</td>
<td>71.4</td>
</tr>
<tr>
<td>- remove $v \odot c_{t-1}$</td>
<td>70.6</td>
<td>71.1</td>
</tr>
<tr>
<td>- remove $v$-scaling</td>
<td>70.3</td>
<td>71.0</td>
</tr>
<tr>
<td>- remove highway</td>
<td>69.4</td>
<td>69.1</td>
</tr>
</tbody>
</table>

Question answering

Tested on SQuAD benchmark using DrQA (Chen et al. 2017) as the model architecture. SRU exhibit 5x speed-up over LSTM and obtains better EM and F1 scores.

Machine translation

Evaluated on WMT English->German dataset. Compared with Transformer by substituting the feed-forward net with SRU.