Recurrence Residual Learning for Sequence Classification

Yiren Wang, University of Illinois at Urbana-Champaign
yiren@illinois.edu

Fei Tian, Microsoft Research
fetia@microsoft.com

1. Motivation

- **LSTM**
  - Gating Mechanism:
    - Controls information flow with parametric gating units
    - Handles gradient vanish problem in vanilla RNN

- **ResNet**
  - Residual Connecting Mechanism:
    - Learns a residual mapping between layers with identity connections
    - Enhances information flow with extra skip connections
    - Effective for optimizing deep architectures

**Can we leverage ResNet to improve LSTM for modeling sequential data?**

**Sequence Classification Tasks**
- Model long sequence
- New input signals at every time step
- Strong info flow over a long time range
- Necessary to "forget" the less critical historical information

2. Our Work

Can we incorporate ResNet to improve LSTM, and leverage the strengths of both gating mechanism in LSTM, and residual connecting mechanism in ResNet?
- Introduce the residual connecting mechanism into the recurrent structure, and propose recurrent residual networks for sequence learning.
- Present in-depth analysis of the strengths and limitations of LSTM and ResNet in respect of sequence learning.
- Propose two novel models that incorporate the strengths of both mechanisms.

3. Models

I. **Recurrent Residual Network (RRN)**
- Leverage residual learning to directly construct recurrent neural network
- Force a direct information flow in different time steps of RNNs by identity (skip) connections

II. **Gated Residual RNN**
- Incorporate ResNet (residual connections) to improve LSTM
- Combine both gating mechanism and residual connecting mechanism

II – 1. Skip-Connected LSTM (SC-LSTM)
- Introduce identity (SC-LSTM-I) and parametric connections (SC-LSTM-P) into standard LSTM to enhance the information flow
- Add skip connections between two LSTM hidden states with a wide range of distance $L$

$$h_{t+L} = \tanh(c_{t+L}) \odot a_{t+L} + ah_t$$

**Can we leverage ResNet to improve LSTM for modeling sequential data?**

II – 2. Hybrid Residual LSTM (HRL)
- Combine the two independent signals propagated by LSTM and ResNet respectively to form a signal
- The final hidden state is the mean pooling on the two representations

$$h_T^{HRL} = \frac{1}{2}(h_T^{LSTM} + h_T^{RNN})$$

4. Experiments

I. **Experiment Settings**
- **Datasets**
  - AG’s News: 34, 211, 4
  - IMDB: 251, 2, 956
  - 20NG: 267, 11, 991
  - P-MNIST: 784, 784, 10

- **Model Size**
  - Non-Hybrid: 256 x 512, 256 x 512
  - 20NG: 256 x 768
  - P-MNIST: 1 x 100
  - Hybrid: 256 x 384, 256 x 384

II. **Experimental Results**

- **Model/Task**
  - AG’s News
  - IMDB
  - 20NG
  - P-MNIST

<table>
<thead>
<tr>
<th>Model/Task</th>
<th>AG’s News</th>
<th>IMDB</th>
<th>20NG</th>
<th>P-MNIST</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM</td>
<td>91.76%</td>
<td>88.88%</td>
<td>79.21%</td>
<td>90.64%</td>
</tr>
<tr>
<td>RRN</td>
<td>91.19%</td>
<td>89.13%</td>
<td>79.76%</td>
<td>88.63%</td>
</tr>
<tr>
<td>SC-LSTM-P</td>
<td>92.01%</td>
<td>90.74%</td>
<td>82.98%</td>
<td>94.46%</td>
</tr>
<tr>
<td>SC-LSTM-I</td>
<td>92.05%</td>
<td>90.67%</td>
<td>81.85%</td>
<td>94.80%</td>
</tr>
<tr>
<td>LSTM+GRU</td>
<td>91.05%</td>
<td>89.23%</td>
<td>80.12%</td>
<td>90.28%</td>
</tr>
<tr>
<td>HRL</td>
<td>91.90%</td>
<td>90.92%</td>
<td>81.73%</td>
<td>90.33%</td>
</tr>
</tbody>
</table>

5. Conclusions

- Direct adaptation of ResNet performs well in sequence classification.
- Residual learning significantly improves LSTM’s performance, particularly in tasks with fairly long sequences.
- Future work: apply residual learning to other sequential tasks, such as language modeling, and RNN-based neural machine translation.