Recurrent Neural Network for Text Classification with Multi-Task Learning

Pengfei Liu, Xipeng Qiu, Xuanjing Huang

Fudan University

May 21, 2016
Background: Neural Sentence Models have achieved impressive performance.
Motivation: How to make our model learn better?

Incorporate more knowledge into model design
- Linguistic characteristics: Recursive NN (Socher et al. 2012); Tree-LSTM (Tai, Socher, and Manning 2015), stack augmented LSTM (Joulin and Mikolov 2015)
- Characteristics of sensory data: Rolling-CNN (Dieleman, De Fauw, and Kavukcuoglu 2016)

Incorporate more knowledge into training data
- Data augmentation
- Multi-task Learning: (Collobert et al. 2011)
Utilizing multi-task learning to jointly learn several tasks with the aim of mutual benefit

- We integrate RNN into the multi-learning framework for text classification and proposed three frameworks with different shared mechanisms.
- We introduce a gating mechanism, which can better control the information passed by the neuron in shared layers.
Recurrent Neural Network for Specific-Task Text Classification

Recurrent Neural Network

\[ h_t = \begin{cases} 
0 & t = 0 \\
 f(h_{t-1}, x_t) & \text{otherwise} 
\end{cases} \] (1)

Figure 1: Recurrent Neural Network for Classification
Long Short-term Memory

\[
\begin{bmatrix}
\tilde{c}_t \\
o_t \\
i_t \\
f_t
\end{bmatrix} = \begin{bmatrix}
tanh \\
\sigma \\
\sigma \\
\sigma
\end{bmatrix} \left( W_p \begin{bmatrix} x_t \\ h_{t-1} \end{bmatrix} + b_p \right),
\]

(2)

\[c_t = \tilde{c}_t \odot i_t + c_{t-1} \odot f_t,\]

(3)

\[h_t = o_t \odot \tanh (c_t),\]

(4)
Three Sharing Models for RNN based Multi-Task Learning

- Model-I: Uniform-Layer Architecture
- Model-II: Coupled-Layer Architecture
- Model-III: Shared-Layer Architecture
Model-I: Uniform-Layer Architecture

\[
\hat{x}_t^{(m)} = x_t^{(m)} \oplus x_t^{(s)},
\]

\[
h_T^{(m)} = LSTM(\hat{x}^{(m)}).
\]
\[
\tilde{c}_t^{(m)} = \tanh \left( W_c^{(m)} x_t + \sum_{i \in \{m,n\}} g^{(i \rightarrow m)} U_c^{(i \rightarrow m)} h_{t-1}^{(i)} \right)
\]

where

\[
g^{(i \rightarrow m)} = \sigma \left( W_g^{(m)} x_t + U_g^{(i)} h_{t-1}^{(i)} \right)
\]
Model-III: Shared-Layer Architecture

\[ \tilde{c}_t^{(m)} = \tanh \left( W_c^{(m)} x_t + g^{(m)} U_c^{(m)} h_{t-1}^{(m)} + g^{(s \rightarrow m)} U_c^{(s)} h_{t}^{(s)} \right), \]  

where

\[ g^{(m)} = \sigma (W_g^{(m)} x_t + U_g^{(m)} h_{t-1}^{(m)}) \]  
\[ g^{(s \rightarrow m)} = \sigma (W_g^{(m)} x_t + U_g^{(s \rightarrow m)} h_{t}^{(s)}) \]
Training and Datasets

Training Details

- Task-specific output
- Pre-training of the shared layer with neural language model
- Fine Tuning

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Type</th>
<th>Train Size</th>
<th>Dev. Size</th>
<th>Test Size</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>SST-1</td>
<td>Sentence</td>
<td>8544</td>
<td>1101</td>
<td>2210</td>
<td>5</td>
</tr>
<tr>
<td>SST-2</td>
<td>Sentence</td>
<td>6920</td>
<td>872</td>
<td>1821</td>
<td>2</td>
</tr>
<tr>
<td>SUBJ</td>
<td>Sentence</td>
<td>9000</td>
<td>-</td>
<td>1000</td>
<td>2</td>
</tr>
<tr>
<td>IMDB</td>
<td>Document</td>
<td>25,000</td>
<td>-</td>
<td>25,000</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 1: Statistics of the four datasets used in this paper.
Experiment Results

<table>
<thead>
<tr>
<th>Model</th>
<th>SST-1</th>
<th>SST-2</th>
<th>SUBJ</th>
<th>IMDB</th>
<th>AvgΔ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Task</td>
<td>45.9</td>
<td>85.8</td>
<td>91.6</td>
<td>88.5</td>
<td>-</td>
</tr>
<tr>
<td>Joint Learning</td>
<td>46.5</td>
<td>86.7</td>
<td>92.0</td>
<td>89.9</td>
<td>+0.8</td>
</tr>
<tr>
<td>+ Fine Tuning</td>
<td>48.5</td>
<td>87.1</td>
<td>93.4</td>
<td>90.8</td>
<td>+2.0</td>
</tr>
</tbody>
</table>

*Table 2: Results of the uniform-layer architecture.*

<table>
<thead>
<tr>
<th>Model</th>
<th>SST-1</th>
<th>SST-2</th>
<th>SUBJ</th>
<th>IMDB</th>
<th>AvgΔ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Task</td>
<td>45.9</td>
<td>85.8</td>
<td>91.6</td>
<td>88.5</td>
<td>-</td>
</tr>
<tr>
<td>SST1-SST2</td>
<td>48.9</td>
<td>87.4</td>
<td>-</td>
<td>-</td>
<td>+2.3</td>
</tr>
<tr>
<td>SST1-SUBJ</td>
<td>46.3</td>
<td>-</td>
<td>92.2</td>
<td>-</td>
<td>+0.5</td>
</tr>
<tr>
<td>SST1-IMDB</td>
<td>46.9</td>
<td>-</td>
<td>-</td>
<td>89.5</td>
<td>+1.0</td>
</tr>
<tr>
<td>SST2-SUBJ</td>
<td>-</td>
<td>86.5</td>
<td>92.5</td>
<td>-</td>
<td>+0.8</td>
</tr>
<tr>
<td>SST2-IMDB</td>
<td>-</td>
<td>86.8</td>
<td>-</td>
<td>89.8</td>
<td>+1.2</td>
</tr>
<tr>
<td>SUBJ-IMDB</td>
<td>-</td>
<td>-</td>
<td>92.7</td>
<td>89.3</td>
<td>+0.9</td>
</tr>
</tbody>
</table>

*Table 3: Results of the coupled-layer architecture.*
## Experiment Results

<table>
<thead>
<tr>
<th>Model</th>
<th>SST-1</th>
<th>SST-2</th>
<th>SUBJ</th>
<th>IMDB</th>
<th>Avg∆</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Task</td>
<td>45.9</td>
<td>85.8</td>
<td>91.6</td>
<td>88.5</td>
<td>-</td>
</tr>
<tr>
<td>Joint Learning</td>
<td>47.1</td>
<td>87.0</td>
<td>92.5</td>
<td>90.7</td>
<td>+1.4</td>
</tr>
<tr>
<td>+ LM</td>
<td>47.9</td>
<td>86.8</td>
<td>93.6</td>
<td>91.0</td>
<td>+1.9</td>
</tr>
<tr>
<td>+ Fine Tuning</td>
<td><strong>49.6</strong></td>
<td><strong>87.9</strong></td>
<td><strong>94.1</strong></td>
<td><strong>91.3</strong></td>
<td><strong>+2.8</strong></td>
</tr>
</tbody>
</table>

**Table 4:** Results of the shared-layer architecture.

<table>
<thead>
<tr>
<th>Model</th>
<th>SST-1</th>
<th>SST-2</th>
<th>SUBJ</th>
<th>IMDB</th>
</tr>
</thead>
<tbody>
<tr>
<td>NBOV</td>
<td>42.4</td>
<td>80.5</td>
<td>91.3</td>
<td>83.62</td>
</tr>
<tr>
<td>RAE</td>
<td>43.2</td>
<td>82.4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MV-RNN</td>
<td>44.4</td>
<td>82.9</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>RNTN</td>
<td>45.7</td>
<td>85.4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DCNN</td>
<td>48.5</td>
<td>86.8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PV</td>
<td>44.6</td>
<td>82.7</td>
<td>90.5</td>
<td><strong>91.7</strong></td>
</tr>
<tr>
<td>Tree-LSTM</td>
<td><strong>51.0</strong></td>
<td><strong>88.0</strong></td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Multi-Task</td>
<td>49.6</td>
<td>87.9</td>
<td><strong>94.1</strong></td>
<td>91.3</td>
</tr>
</tbody>
</table>
Figure 3: (a) The change of the predicted sentiment score at different time steps. Y-axis represents the sentiment score, while X-axis represents the input words in chronological order. The red horizontal line gives a border between the positive and negative sentiments. (b) Visualization of the global gate’s ($g^{(s)}$) activation.
we introduce three RNN based architectures to model text sequence with multi-task learning.

Experimental results show that our models can improve the performances of a group of related tasks by exploring common features.
Thank you!
References

Ammar, Waleed et al. (2016). “Many languages, one parser”. In: CoRR, abs/1602.01595.


Dong, Daxiang et al. (2015). “Multi-Task Learning for Multiple Language Translation”. In: Proceedings of the ACL.
