Gated Feedback
Recurrent Neural Networks

Junyoung Chung, Caglar Gulchre, Kyunghyun Cho and Yoshua Bengio

Presenter: Junyoung Chung
Challenges in Sequence Modelling

• Sequences contain hierarchical structures like images
  • Characters / phonemes change fast
  • Topics of a corpus change slowly
  • Genders of an utterance do not change

• RNN should encode multiple time-scale temporal dependencies
  • RNNs process a sequence in a step by step manner
Multi-Scale RNNs

- A Hierarchical RNN [Hihi and Bengio, 1995]
  - Sequences have hierarchical temporal dependencies
  - Multiple levels of states operating at different time-scales, e.g. power of 2
  - Slow level (coarse time-scale) can affect fast level (fine time-scale)
Multi-Scale RNNs

- A Clockwork RNN [Koutník et al., 2014]
  
  - Multiple modules operating with pre-defined clock rates
  
  - $i$-th module updates its states when $t \mod 2^{i-1} = 0$
  
  - Slow module (coarse time-scale) can affect fast module (fine time-scale)
Multi-Scale RNNs

• Why are multi-scale representations useful?
  
  - RNNs use richer representation to draw desired output at each timestep
  - Skip connections [Graves, 2013] from all hidden layers to output layer
Multi-Scale RNNs

• Why are top-down connections helpful?

  - Dynamics of coarser time-scales summarize global structure
  - Dynamics of coarser time-scales can help inference of finer time-scales
  - We do not precisely know the update rates
Gated Feedback RNNs

• Interesting architectural modifications to RNNs
  
  • Can we adaptively learn the update rate of each recurrent layer?
  
  • Can multi-scale representations used to compute the hidden states of each recurrent layer?
Gated Feedback RNNs

- Interesting architectural modifications to RNNs
  - Can we adaptively learn the update rate of each recurrent layer?  **Yes**
  - Can multi-scale representations used to compute the hidden states of each recurrent layer?  **Yes**
Recurrent Activation Functions

- Recurrent activation functions determine the power of RNNs
  - Deterministic dynamics
  - Gated activation functions are more powerful e.g., GRU, LSTM
Global Reset Gates

- **Global reset gates** adaptively control the connectivity patterns

  - Inputs: current state of layer below, concatenation of all previous states

  \[ g^{i \rightarrow j} = \text{sigm} \left( w^{i \rightarrow j}_{g} h^{j-1}_t + u^{i \rightarrow j}_{g} h^{*}_{t-1} \right) \]

  - Number of global reset gates should be number of layers\(^2\)

<table>
<thead>
<tr>
<th>Standard RNNs</th>
<th>Gated Feedback RNNs</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \tanh )</td>
<td>( \tanh \left( W^j h^{j-1}<em>t + \sum</em>{i=1}^{L} g^{i \rightarrow j} U^{i \rightarrow j} h^i_{t-1} \right) )</td>
</tr>
<tr>
<td>( \text{GRU} )</td>
<td>( \text{GRU} )</td>
</tr>
<tr>
<td>( \tilde{c}^j_t = \tanh(W h^{j-1}<em>t + U h^{j}</em>{t-1}) )</td>
<td>( \tilde{c}^j_t = \tanh \left( W^j h^{j-1}<em>t + \sum</em>{i=1}^{L} g^{i \rightarrow j} U^{i \rightarrow j} h^i_{t-1} \right) )</td>
</tr>
<tr>
<td>( \tilde{h}^j_t = \tanh(W h^{j-1}<em>t + r \odot U h^{j}</em>{t-1}) )</td>
<td>( \tilde{h}^j_t = \tanh \left( W^j h^{j-1}<em>t + r^j_t \odot \sum</em>{i=1}^{L} g^{i \rightarrow j} U^{i \rightarrow j} h^i_{t-1} \right) )</td>
</tr>
<tr>
<td>( \text{LSTM} )</td>
<td>( \text{LSTM} )</td>
</tr>
<tr>
<td>( \tilde{c}^j_t = \tanh(W h^{j-1}<em>t + U h^{j}</em>{t-1}) )</td>
<td>( \tilde{c}^j_t = \tanh \left( W^j h^{j-1}<em>t + \sum</em>{i=1}^{L} g^{i \rightarrow j} U^{i \rightarrow j} h^i_{t-1} \right) )</td>
</tr>
</tbody>
</table>
Experiments

• Tasks
  - Character-level language modelling
  - Python program evaluation

• Models
  - Three RNN architectures
    - Single layer RNN, Stacked RNNs, Gated Feedback RNNs
  - Three activation functions
    - Affine transformation + $\tanh$, GRU, LSTM
Character-Level Language Modelling

- Measure the bits-per-character \( E[- \log_2 P(x_{t+1} | h_t)] \)

- Constrained model size

<table>
<thead>
<tr>
<th></th>
<th>tanh</th>
<th>GRU</th>
<th>LSTM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-layer</td>
<td>1.937</td>
<td>1.883</td>
<td>1.887</td>
</tr>
<tr>
<td>Stacked</td>
<td>1.892</td>
<td>1.871</td>
<td>1.868</td>
</tr>
<tr>
<td>Gated Feedback</td>
<td>1.949</td>
<td>1.855</td>
<td>1.842</td>
</tr>
<tr>
<td>Gated Feedback L</td>
<td>–</td>
<td>1.813</td>
<td>1.789</td>
</tr>
<tr>
<td>Feedback*</td>
<td>–</td>
<td>–</td>
<td>1.854</td>
</tr>
</tbody>
</table>

- Unconstrained model size

<table>
<thead>
<tr>
<th></th>
<th>MRNN</th>
<th>Stacked LSTM</th>
<th>GF-LSTM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.60</td>
<td>1.67</td>
<td>1.58</td>
</tr>
</tbody>
</table>

- MRNN [Sutskever et al., 2011], 86 characters, XML tags removed
- Stacked LSTM [Graves, 2013], 205 characters
Character-Level Language Modelling

- Predicting XML tags with stacked LSTM and GF-LSTM
Character-Level Language Modelling

- Generated samples given various kinds of test seed snippets

Istanbul is located in north-western [[Turkey]] and south-eastern Europe within the [[Marmara Region]] on a total area of 600km² and wards by merging the airport and carry [[Discount of Cuba|Year of Empire|Temastique]], "quote;

*[[http://indianajones.wikicities.com The Indiana Jones Wiki]]
*[[http://www.theindyexperience.com The Indy Experience]]
*[[http://www.indygear.org/ Eugenics]],
which presumably includes an inside Liberals or Civilian leader JRD C President. Two large faculty alien colonies led by Thomas

### **[[Methodism|Methodist]]** &amp;ndash; 10%
### **[[Lutheranism|Lutheran]]** &amp;ndash; 6%
### **[[Church of Christ]]** &amp;ndash; 6%
### **[[Reformed Egyptian|Egyptian]]** &amp;ndash; 11%
### **[[Syriac]]** – Member of Augusta (cultural)

In Europe, Paris was the center stage for the [[French Revolution]], and it became an important centre of finance, commerce, and funding important.
The markets for the Union also include the invention of money lower risks among a possible sacrifice and electrotex and their chemical

Google was founded by [[Larry Page]] and [[Sergey Brin]] while they were [[Doctor of Philosophy|Ph.D.]] students at [[Stanford University]]. Together they own about 14,000 files

fimber[266] [[January 15|15]] [[January 12|12]] [[January 1]].
Leading the democratisation of the Czech Republic in 1948,
Python Program Evaluation

• An interesting task to show if an RNN can understand programs

  • The goal is to predict the correct return value of a given Python script

  • Addition, multiplication, substraction, for-loop, variable assignment, logical comparison and if-else statement

  • Can artificially control the difficulty of each example

  • A training example looks like

```
Input:
j=8584
for x in range(8):
    j+=920
b=(1500+j)
print((b+7567))
```

```
Target: 25011.
```

Image credit: Learning to Execute [Zaremba and Sutskever, 2014]
Python Program Evaluation

- Python program evaluation results

GRU

Stacked RNNs

Gated Feedback RNNs

△ Gated Feedback—Stacked
Conclusion

• Gated-feedback architecture is helpful
  - Gated-feedback RNNs converge into better optimal point (even faster)
  - Gated-feedback RNNs take advantage when the tasks become harder
  - In all of our experiments, GF-LSTM scored the best results

• However, gated-feedback architecture performs worse when it is combined with $\tanh$
Thank you