RECURRENT CONVOLUTIONAL NEURAL NETWORKS FOR SPEECH PROCESSING

Yue Zhao, Xingyu Jin
Department of Electronic Engineering, TNList,
Tsinghua University, Beijing, 100084, China

Xiaolin Hu
Department of Computer Science and Technology,
TNList, Tsinghua University, Beijing, 100084, China

**MOTIVATION**
- Existing CNN and RNN have specific disadvantages.
  - CNN has not exhibited significant improvement in speech processing.
  - RNN is expected to function well in modeling sequential, but is harder to train efficiently.
- A new architecture of Recurrent Convolutional Neural Network (RCNN) [1, 2] works well in object recognition and scene labeling.
- In view of the embedded RNN structure, RCNN is expected to function well in modeling speech, a typical temporally sequential data.

**FORMULATION**
- Conventional RNN: (recursing less term)
  \[ h(t) = \sigma(W_gh h(t-1) + W_gx x(t)) \]
  \( x(t) \) : feed-forward input, \( h(t) \) : hidden state at time \( t \)
- Recurrent Convolutional Layer (RCL): 
  \[ h(t)(i) = \sum_{j} W_t(i,j)' x(j) \] 
  \[ + \sum_{j} W_s(i,j)' h(t-1)(j) \] 
  \( W_t^{(k)} \) : feedforward kernel, \( W_s^{(k)} \) : recurrent kernel
- Nonlinearity \( \sigma(x) = f_{max}(g(x)) \) is realized by Rectified Linear (ReLU) \( g(x) = \max(x, 0) \) and batch normalization \( f_{max}(x; \gamma, \beta) \).
- “time step” \( t \) in RCL: a RCL processes information from neighboring time slots and frequency banks at each iteration.

**ILLUSTRATION**
Illustration of a single RCL and its unfolded version with \( T=3 \).

**RESULTS**
- Phoneme recognition on TIMIT
  - Unfolding more times yields lower PER but there is a limit.
  - Outperform most ANN/HMM models. Competitive to existing methods. (more in paper)

<table>
<thead>
<tr>
<th>Training system</th>
<th>4-layer MLP</th>
<th>5-layer MLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>PER (dev set)</td>
<td>19.9%</td>
<td>18.4%</td>
</tr>
<tr>
<td>PER (core test set)</td>
<td>22.0%</td>
<td>20.5%</td>
</tr>
</tbody>
</table>

CL + pooling + 3-layer MLP
RCL (T=1) + pooling + 3-layer MLP
RCL (T=2) + pooling + 3-layer MLP
RCL (T=2) + CL + 9-layer MLP
3-layer LSTM + HMM [3]

- The speed of RCNN is faster than LSTM module, both when training and decoding.

**CONCLUSIONS**
- Propose to use RCNN originally from computer vision to speech processing.
- RCNN achieves competitive results with existing models. Also, it runs faster than LSTM networks.
- Inspire more generic and efficient cross-modal deep learning models in the future.

**REFERENCES**

**SOURCE CODES**
The source codes can be downloaded at: https://github.com/zhaoyue-thuzhaoyue/RecurrentConvNet

Contact: Yue Zhao, thuzhaoyue@gmail.com