Densely Connected Graph Convolutional Networks for Graph-to-Sequence Learning

Zhijiang Guo

Joint work with Yan Zhang, Zhiyang Teng, Wei Lu
Graph-to-Sequence Learning

AMR-to-Text Generation

Syntax-Based Machine Translation
AMR-to-Text Generation

You guys know what I mean.
Sequence Encoder

BiLSTM

know :arg1 thing :arg0 mean :arg0 i :arg0 you :mod guy

(Konstas et al. 2017)
Recurrent Graph Encoder

Graph State LSTM (Song et al., 2018)
Gated Graph Neural Networks (Beck et al., 2018)
Empirically, the best performance of GCNs is achieved with a **2-layer** model (Li et al., 2018, Xu et al., 2018)
GCNs

first convolutional layer captures **first-order proximity** (immediate neighbors) information
GCNs

second convolutional layer captures second-order proximity information
Convolutional Graph Encoder

(Bastings et al., 2017) (Damonte and Cohen, 2019)
Motivation

Is it possible to build a more expressive GCN model to learn a better graph representation without relying on additional LSTM?

Densely Connected Graph Convolutional Networks (DCGCNs)
Dense Connectivity

one layer takes inputs from all preceding layers rather than the previous layer only (Huang et al., 2017)
Densely Connected GCNs

Stack Identical Blocks

- Densely Connected Sub-Block
- Linear Combination Layer
Both sub-blocks are densely connected graph convolutional layers with different numbers (m and n) of layers.
Densely Connected Sub-Block

Sub-blocks with different number of layers capture structural information at **different abstract levels**, similar to different filters.
Densely Connected Sub-Block

For parameter efficiency, the output dimension of each layer in the sub-block is designed to be small.
Densely Connected Sub-Block

Input dimension: 300
Sub-block layers: 3

Hidden dimension of each layer:
100 = 300 / 3 (proportional to #layers)

Output dimension: 300
(concatenate output from all 3 layers)
Linear Combination Layer

This layer assigns **different weights** to outputs of different layers. Initial inputs of the sub-block are also incorporate by the residual connection.
Graph-to-Sequence Model

Linear Combination

Attention Mechanism

DCGCN Block

DCGCN Block

DCGCN Block

LSTM Layer

LSTM Layer

Positional Encoding

Input Embedding

Output Embedding

Encoder

Decoder
Experiments

AMR-to-Text Generation

- AMR 2015
- AMR 2017

Syntax-Based Machine Translation

- English-Czech (WMT 16)
- English-German (WMT 16)
### Data Statistics

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMR 2015</td>
<td>16,833</td>
<td>1,368</td>
<td>1,371</td>
</tr>
<tr>
<td>AMR 2017</td>
<td>36,521</td>
<td>1,368</td>
<td>1,371</td>
</tr>
<tr>
<td>En-Cs</td>
<td>181,112</td>
<td>2,656</td>
<td>2,999</td>
</tr>
<tr>
<td>En-De</td>
<td>226,822</td>
<td>2,169</td>
<td>2,999</td>
</tr>
</tbody>
</table>
**AMR 2015**

**Sequential Encoder:** LSTM (Konstas et al., 2017)

**Graph Encoder:** GS LSTM (Song et al., 2018)
- GCN + LSTM (Damonte and Cohen, 2019)

<table>
<thead>
<tr>
<th>Model</th>
<th>External Data</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM</td>
<td>No</td>
<td>22.0</td>
</tr>
<tr>
<td>GS LSTM</td>
<td>No</td>
<td>23.3</td>
</tr>
<tr>
<td>GCN + LSTM</td>
<td>No</td>
<td>24.4</td>
</tr>
<tr>
<td>DCGCN</td>
<td>No</td>
<td>25.7</td>
</tr>
</tbody>
</table>
## AMR 2015

### Using External Training Data (0.2M)

<table>
<thead>
<tr>
<th>Model</th>
<th>External Data</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM</td>
<td>0.2M</td>
<td>27.4</td>
</tr>
<tr>
<td>GS LSTM</td>
<td>0.2M</td>
<td>28.2</td>
</tr>
<tr>
<td>DCGCN</td>
<td>0.1M</td>
<td>29.0</td>
</tr>
<tr>
<td>DCGCN</td>
<td>0.2M</td>
<td>31.6</td>
</tr>
</tbody>
</table>
## AMR 2015

Using External Training Data (0.3M)

<table>
<thead>
<tr>
<th>Model</th>
<th>External Data</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM</td>
<td>2M</td>
<td>32.3</td>
</tr>
<tr>
<td>LSTM</td>
<td>20M</td>
<td>33.8</td>
</tr>
<tr>
<td>GS LSTM</td>
<td>2M</td>
<td>33.6</td>
</tr>
<tr>
<td>DCGCN (Single)</td>
<td>0.3M</td>
<td>33.2</td>
</tr>
<tr>
<td>DCGCN (Ensemble)</td>
<td>0.3M</td>
<td>35.3</td>
</tr>
</tbody>
</table>
#AMR 2017 (Single)

**Sequential Encoder:** LSTM (Beck et al., 2017)

**Graph Encoder:** GGNNs (Beck et al., 2018)

**GCN + LSTM** (Damonte and Cohen, 2019)

<table>
<thead>
<tr>
<th>Model</th>
<th>#Parameters</th>
<th>BLEU</th>
<th>CHRF++</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM</td>
<td>28.4M</td>
<td>21.7</td>
<td>49.1</td>
</tr>
<tr>
<td>GGNNs</td>
<td>28.3M</td>
<td>23.3</td>
<td>50.4</td>
</tr>
<tr>
<td>GCN + LSTM</td>
<td>N/A</td>
<td>24.5</td>
<td>N/A</td>
</tr>
<tr>
<td>DCGCN</td>
<td>18.5M</td>
<td>27.6</td>
<td>57.3</td>
</tr>
</tbody>
</table>
AMR 2017 (Ensemble)

Sequential Encoder: LSTM (Beck et al., 2017)
Graph Encoder: GGNNs (Beck et al., 2018)

<table>
<thead>
<tr>
<th>Model</th>
<th>#Parameters</th>
<th>BLEU</th>
<th>CHRF++</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM</td>
<td>142.0M</td>
<td>26.6</td>
<td>52.5</td>
</tr>
<tr>
<td>GGNNs</td>
<td>141.0M</td>
<td>27.5</td>
<td>53.5</td>
</tr>
<tr>
<td>DCGCN</td>
<td>92.5M</td>
<td>30.4</td>
<td>59.6</td>
</tr>
</tbody>
</table>
### English-German

**Sequential Encoder:** LSTM (Konstas et al., 2017)

**Graph Encoder:** GGNNs (Beck et al., 2018)

BoW/CNN/RNN + GCN (Bastings et al., 2017)

<table>
<thead>
<tr>
<th>Model</th>
<th>Type</th>
<th>#Param</th>
<th>BLEU</th>
<th>CHRF++</th>
</tr>
</thead>
<tbody>
<tr>
<td>BoW + GCN</td>
<td>Single</td>
<td>N/A</td>
<td>12.2</td>
<td>N/A</td>
</tr>
<tr>
<td>CNN + GCN</td>
<td>Single</td>
<td>N/A</td>
<td>13.7</td>
<td>N/A</td>
</tr>
<tr>
<td>BiRNN + GCN</td>
<td>Single</td>
<td>N/A</td>
<td>16.1</td>
<td>N/A</td>
</tr>
<tr>
<td>Seq2Seq</td>
<td>Single</td>
<td>41.4M</td>
<td>15.5</td>
<td>40.8</td>
</tr>
<tr>
<td>GGNNs</td>
<td>Single</td>
<td>41.2M</td>
<td>16.7</td>
<td>42.4</td>
</tr>
<tr>
<td>Our DCGCN</td>
<td>Single</td>
<td>29.7M</td>
<td>19.0</td>
<td>44.1</td>
</tr>
</tbody>
</table>
**English-Gzech**

**Sequential Encoder:** LSTM (Konstas et al., 2017)

**Graph Encoder:** GGNNs (Beck et al., 2018)

BoW/CNN/RNN + GCN (Bastings et al., 2017)

<table>
<thead>
<tr>
<th>Model</th>
<th>Type</th>
<th>#Param</th>
<th>BLEU</th>
<th>CHRF++</th>
</tr>
</thead>
<tbody>
<tr>
<td>BoW + GCN</td>
<td>Single</td>
<td>N/A</td>
<td>7.5</td>
<td>N/A</td>
</tr>
<tr>
<td>CNN + GCN</td>
<td>Single</td>
<td>N/A</td>
<td>8.7</td>
<td>N/A</td>
</tr>
<tr>
<td>BiRNN + GCN</td>
<td>Single</td>
<td>N/A</td>
<td>9.6</td>
<td>N/A</td>
</tr>
<tr>
<td>Seq2Seq</td>
<td>Single</td>
<td>41.4M</td>
<td>8.9</td>
<td>33.8</td>
</tr>
<tr>
<td>GGNNs</td>
<td>Single</td>
<td>41.2M</td>
<td>9.8</td>
<td>33.3</td>
</tr>
<tr>
<td>Our DCGCN</td>
<td>Single</td>
<td>29.7M</td>
<td>12.1</td>
<td>37.1</td>
</tr>
</tbody>
</table>
# Density of Connection

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCGCN</td>
<td>25.5</td>
</tr>
<tr>
<td>- {4} dense block</td>
<td>24.8</td>
</tr>
<tr>
<td>- {3, 4} dense block</td>
<td>23.8</td>
</tr>
<tr>
<td>- {2, 3, 4} dense blocks</td>
<td>23.2</td>
</tr>
</tbody>
</table>
## Ablation Test

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCGCN</td>
<td>25.5</td>
</tr>
<tr>
<td>- Global Node (GN)</td>
<td>24.2</td>
</tr>
<tr>
<td>- Linear Combination (LC)</td>
<td>23.7</td>
</tr>
<tr>
<td>- GN, LC</td>
<td>22.9</td>
</tr>
</tbody>
</table>
Conclusion

- DCGCNs allow the encoder to better capture the rich structural information of a graph, especially when it is large.

- Future: investigate how other NLP applications can potentially benefit from our proposed approach.
Thank You

Code available:

http://www.statnlp.org/research/machine-learning