DL4NLP: Challenges and Future Directions

Xipeng Qiu
xpqiu@fudan.edu.cn
http://nlp.fudan.edu.cn/~xpqiu

Fudan University

November 14, 2015
CCL 2015
Guangzhou, China
Outline

1 Neural Models for NLP
2 DL4NLP at Fudan NLP Lab
   - Our Focused Problem: Feature Composition
   - Convolutional Neural Tensor Network
   - Recursive Neural Network for Dependency Parse Tree
   - Gated Recursive Neural Network
   - Multi-Timescale LSTM
3 Future Directions
   - Memory Mechanism
   - Attention Mechanism
   - Novel Applications
General Neural Architectures for NLP

1. represent the words/features with dense vectors (embeddings) by lookup table;
2. concatenate the vectors;

from [Collobert et al., 2011]
## Difference with the traditional methods

<table>
<thead>
<tr>
<th>Features</th>
<th>Traditional methods</th>
<th>Neural methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Features</td>
<td>Discrete Vector (One-hot Representation)</td>
<td>Dense Vector (Distributed Representation)</td>
</tr>
<tr>
<td></td>
<td>High-dimension</td>
<td>Low-dimension</td>
</tr>
<tr>
<td>Classifier</td>
<td>Linear</td>
<td>Non-Linear</td>
</tr>
</tbody>
</table>
General Neural Architectures for NLP

- **Word Level**
  - NNLM
  - C&W
  - CBOW & Skip-Gram

- **Sentence Level**
  - NBOW
  - **Sequence Models**: Recurrent NN, LSTM, Paragraph Vector
  - **Topological Models**: Recursive NN,
  - **Convolutional Models**: DCNN

- **Document Level**
  - NBOW
  - **Hierarchical Models** two-level CNN
  - **Sequence Models** LSTM, Paragraph Vector
Not “Really” Deep Learning in NLP

- Most of the neural models is very shallow in NLP.
- The major benefit is introducing dense representation.
- The feature composition is also quite simple.
  - Concatenation
  - Sum/Average
  - Bilinear model
Quite Simple Feature Composition

Given two embeddings $a$ and $b$,

1. how to calculate their similarity/relevance/relation?
   - Concatenation
     $$a \oplus b \rightarrow \text{ANN} \rightarrow \text{output}$$
   - Bilinear
     $$a^T Mb \rightarrow \text{output}$$

2. how to use them in classification task?
   - Concatenation
     $$a \oplus b \rightarrow \text{ANN} \rightarrow \text{output}$$
   - Sum/Average
     $$a + b \rightarrow \text{ANN} \rightarrow \text{output}$$
Problem

How to enhance the neural model without increasing the network depth?
Convolutional Neural Network (CNN)

Key steps
- Convolution
- (optional) Folding
- Pooling

Various models
- DCNN (k-max pooling) [Kalchbrenner et al., 2014]
- CNN (binary pooling) [Hu et al., 2014]
- ...
Convolutional Neural Tensor Network for Text Matching

[Qiu and Huang, 2015]

Architecture of Convolutional Neural Tensor Network
Recursive Neural Network (RecNN) [Socher et al., 2013]

Topological models compose the sentence representation following a given topological structure over the words.

Given a labeled binary parse tree, \(((p_2 \rightarrow ap_1), (p_1 \rightarrow bc))\), the node representations are computed by

\[ p_1 = f(W \begin{bmatrix} b \\ c \end{bmatrix}), \]

\[ p_2 = f(W \begin{bmatrix} a \\ p_1 \end{bmatrix}). \]
A variant of RecNN for Dependency Parse Tree [Zhu et al., 2015]

Recursive neural network can only process the binary combination and is not suitable for dependency parsing.

Recursive Convolutional Neural Network

- introducing the convolution and pooling layers;
- modeling the complicated interactions of the head word and its children.
Gated Recursive Neural Network [Chen et al., 2015a]

- DAG based Recursive Neural Network
- Gating mechanism

An relative complicated solution

GRNN models the complicated combinations of the features, which selects and preserves the useful combinations via reset and update gates.

A similar model: AdaSent [Zhao et al., 2015]
Our Focused Problem: Feature Composition
Convolutional Neural Tensor Network
Recursive Neural Network for Dependency Parse Tree
Gated Recursive Neural Network
Multi-Timescale LSTM

Two Gates
- reset gate
- update gate

- Chinese Word Segmentation [Chen et al., 2015a]
- Dependency Parsing [Chen et al., 2015c]
- Sentence Modeling [Chen et al., 2015b]
Unfolded LSTM for Text Classification

Drawback: long-term dependencies need to be transmitted one-by-one along the sequence.
Multi-Timescale LSTM

Figure: Two feedback strategies of our model. The dashed line shows the feedback connection, and the solid link shows the connection at current time.

from [Liu et al., 2015]
Unfolded Multi-Timescale LSTM with Fast-to-Slow Feedback Strategy

from [Liu et al., 2015]
LSTM for Sentiment Analysis

<s> Is this progress ? </s>
LSTM
MT-LSTM

<s> He ’d create a movie better than this . </s>
LSTM
MT-LSTM

<s> It ’s not exactly a gourmet meal but the fare is fair , even coming from the drive . </s>
LSTM
MT-LSTM
### Memory Mechanism

What differences among the various models from memory view?

<table>
<thead>
<tr>
<th>Models</th>
<th>Short-term</th>
<th>long-term</th>
<th>Global</th>
<th>External</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRN</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>LSTM/GRU</td>
<td>Yes</td>
<td>No</td>
<td>Maybe</td>
<td>No</td>
</tr>
<tr>
<td>PV</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>NTM/DMN</td>
<td>Maybe</td>
<td>Maybe</td>
<td>Maybe</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Neural Models as Components

- Component models could be more complex than main model.
- More attention mechanisms?
Novel Applications

- Abstractive Summarization
- Text Generation
- Integration of Syntax, Semantics and Knowledge
- ...


References II


References III
