Convolutional Neural Networks for Text Classification

Sebastian Sierra

MindLab Research Group

July 1, 2016
Outline

1. What is a Convolution?
2. What are Convolutional Neural Networks?
3. CNN for NLP
4. CNN hyperparameters
5. Example: The Model
6. Bibliography
Outline

1. What is a Convolution?

2. What are Convolutional Neural Networks?

3. CNN for NLP

4. CNN hyperparameters

5. Example: The Model

6. Bibliography
What is a Convolution?

- Convolutions are great for extracting features from Images.
- Convolutional Neural Networks (CNN) are biologically-inspired variants of MLPs
Green: Input, Yellow: Convolutional Filter, Red: Output
Green: Input, Yellow: Convolutional Filter, Red: Output
Green: Input, Yellow: Convolutional Filter, Red: Output
Green: Input, Yellow: Convolutional Filter, Red: Output
Green: Input, Yellow: Convolutional Filter, Red: Output
Green: Input, Yellow: Convolutional Filter, Red: Output
Green: Input, Yellow: Convolutional Filter, Red: Output
Green: Input, Yellow: Convolutional Filter, Red: Output
Green: Input, Yellow: Convolutional Filter, Red: Output
Outline

1. What is a Convolution?

2. What are Convolutional Neural Networks?

3. CNN for NLP

4. CNN hyperparameters

5. Example: The Model

6. Bibliography
Convolutional Neural Network

Figure 1: Close up of Convolutional Neural Network
Convolutional Neural Network

- CNNs are networks composed of several layers of convolutions with nonlinear activation functions like ReLU or tanh applied to the results.
- Traditional Layers are fully connected, instead CNN use local connections.
- Each layer applies different filters (thousands) like the ones showed above, and combines their results.
Properties of Convolutional Neural Networks

- Local Invariance
- Compositionality
Do they make sense in NLP?
Perhaps Recurrent Neural Networks would make more sense trying to learn patterns extracted from a text sequence. They are not cognitively or linguistically plausible.

Advantage
There are fast GPU implementations for CNNs
An example of how CNN work

3-gram filter

<table>
<thead>
<tr>
<th>0</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>G</th>
<th>T</th>
<th>C</th>
<th>A</th>
<th>A</th>
<th>C</th>
<th>A</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
An example of how CNN work
An example of how CNN work
An example of how CNN work

3-gram filter

```
0 0 0
0 0 1
1 0 0
0 1 0
```

```
3 0 0 1 0 0
```

```
0 0 0 1 1 0 1 0
0 0 1 0 0 1 0 0
1 0 0 0 0 0 0 0
0 1 0 0 0 0 0 1
```

G T C A A C A T
An example of how CNN work
An example of how CNN work
An example of how CNN work

2-gram filter

0 1
1 0
0 0
0 1

0 0 0 1 1 0 1 0
0 0 1 0 0 1 0 0
1 0 0 0 0 0 0 0
0 1 0 0 0 0 0 1

G T C A A C A T
An example of how CNN work
An example of how CNN work

Several filters

<table>
<thead>
<tr>
<th>G</th>
<th>T</th>
<th>C</th>
<th>A</th>
<th>A</th>
<th>C</th>
<th>A</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
An example of how CNN work

Several filters

They are fed to the FC layer.

We expect to capture high level features at this layer

hidden_dims: 150

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

G  T  C  A  A  C  A  T
Char-CNN

Let’s try this model
Char-CNN

Let’s try this model

\( C \in \mathbb{R}^{d \times l} \) : Matrix representation of a sequence of length \( l \) (140, 300, ?).

\( H \in \mathbb{R}^{d \times w} \) : Convolutional filter matrix where,

- \( d \) : Dimensionality of character embeddings (used 30)
- \( w \) : Width of convolution filter (3, 4, 5)
A simple architecture
Steps for applying a CNN

1. Apply a convolution between $C$ and $H$ to obtain a vector $f \in \mathbb{R}^{l-w+1}$

$$f[i] = \langle C[*, i : i + w - 1], H \rangle$$

$\langle A, B \rangle$ is the Frobenius inner product. $Tr(AB^T)$

2. This vector $f$ is also known as feature map.

3. Take the maximum value over time as the feature that represents filter $H$. (K-max pooling)

$$\hat{f} = \text{relu}(\max_i \{f[i]\} + b)$$

4. Then we do the same for all $m$ filters.

$$z = [\hat{f}_i, ..., \hat{f}_m]$$
Intuition behind

- Why \textit{ReLU} and not \textit{tanh}?
- Should I use multiple filter weights \( H \)?
- Should I use variable filter widths \( w \)?
- Can I add another channel as in Computer Vision domain?
- Is Max-Pooling capturing the most important activation?
- Would they capture morphological relations?
Outline

1. What is a Convolution?
2. What are Convolutional Neural Networks?
3. CNN for NLP
4. CNN hyperparameters
5. Example: The Model
6. Bibliography
Regularization tricks

- Use dropout (Gradients are only backpropagated through certain inputs of $z$).
- Constrain $L_2$ norms of weight vectors of each class (rows in Softmax matrix $W^{(S)}$) to a fixed number: If $\|W_c^{(S)}\| > s$, the rescale it so $\|W_c^{(S)}\| = s$.
- Early Stopping
Previous works: NLP from scratch (Collobert et al. 2011).
Sentence or paragraph modelling using words as input (Kim 2014; Kalchbrenner, Grefenstette, and Blunsom 2014; Johnson and T. Zhang 2015a; Johnson and T. Zhang 2015b).
Text classification using characters as input (Kim et al. 2016; X. Zhang, Zhao, and LeCun 2015)
Outline

1. What is a Convolution?
2. What are Convolutional Neural Networks?
3. CNN for NLP
4. CNN hyperparameters
5. Example: The Model
6. Bibliography
Outline

1. What is a Convolution?
2. What are Convolutional Neural Networks?
3. CNN for NLP
4. CNN hyperparameters
5. Example: The Model
6. Bibliography


