Distributed Representations of Sentences and Documents

Quoc Le and Tomas Mikolov

(ICML 2014)
Discussion by: Chunyuan Li

April 17, 2015
Outline

1. Word Vector
   - Background
   - Neural Language Model
   - Continuous Bag-of-Words
   - Skip-gram Model

2. Paragraph Vector
   - Distributed Memory Model of Paragraph Vectors
   - Distributed Bag of Words of Paragraph Vector

3. Experiments
   - Sentiment Analysis
   - Information Retrieval
Background in text representation

- One-hot representation/One-of-N coding
- Bag-of-words
- N-gram model
Neural Language Model

- A mapping $C$ from any element $i$ of $V$ to a real vector $C(i)$. It represents the **distributed feature vectors**.
- Learning in **context**.

  “The **cat** is walking in the **bedroom**”

- Maximize the average (regularized) log-likelihood

$$L = \frac{1}{T} \sum_t \log f(w_t, w_{t-1}, \ldots, w_{t-(n-1)}; \theta)$$

*A neural probabilistic language model* (Bengio et al. JMLR 2003)
Neural Language Model

- A conditional probability distribution over words in $V$ for the next word $w_t$

$$p(w_t | w_{t-1}, \cdots, w_{t-n+1}) = \frac{\exp(y_{w_t})}{\sum_i \exp(y_i)}$$

where

$$y = b + Wx + U \tanh(d + Hx)$$

$$x = (C(w_{t-1}), C(w_{t-2}), \cdots, C(w_{t-(n-1)}))$$

$$\theta = (b, d, W, U, H, C)$$

red: model parameters, green: vector representation
Continous Bag-of-Words (Mikolov et al, 2013)

- Predict the current word based on the context
- The nonlinear hidden layer is removed
- \( y = b + Wx \)
- \( \theta = (b, W, C) \)

Efficient estimation of word representations in vector space (Mikolov et al, 2013)
Skip-gram Model

- Predict the surrounding words

\[ f = \prod_{-\ell \leq j \leq \ell, j \neq 0} p(w_{t+j} | w_t) \]

where

\[ p(w_{t+j} | w_t) = \frac{\exp(y_{w_t+j}^\top y_{w_t})}{\sum_i \exp(y_i^\top y_{w_t})} \]

- \( y_i = C(w_i) \)
- \( \theta = C \)

*Distributed representations of words and phrases and their compositionality* (Mikolov et al, NIPS 2013)
Word Vector - Linguistic Regularities

- One can do nearest neighbor search around result of vector operation

"King – man + woman" and obtain "Queen"

*Mikolov et al., 2013*
Distributed Memory Model of Paragraph Vectors (PV-DM)

- $D$: paragraph vectors; $W$: word vectors
- $x$ is constructed from $W$ and $D$
- It acts as a memory that remembers what is missing from the current context
- One paragraph vector is only shared across all contexts generated from the same paragraph; The word vector is shared across paragraphs.
Distributed Bag of Words of Paragraph Vector (PV-DBOW)

- In practice
  1. sample a text window
  2. sample a random word from the text window
  3. form a classification task given the Paragraph Vector.

- PV-DM alone usually works well for most tasks. The final paragraph vector is a combination of two vectors.
Experiment I: Sentiment Analysis

Sentiment analysis
- Stanford sentiment treebank dataset (Socher et al., 2013b)
- IMDB dataset (Maas et al., 2011)

Evaluation
- Fine-grained: \{Very Negative, Negative, Neutral, Positive, Very Positive\}
- Coarse-grained: \{Negative, Positive\}

Methods to compare
- Bag-of-Words
- Word Vector Averaging (Socher et al., 2013b)
- Recursive Neural Network (Socher et al., 2011)
- Martix Vector-RNN (Socher et al., 2012)
- Recursive Neural Tensor Network (Socher et al., 2013)
Recursive Neural Network (RNN)

- Each node is attached 3 items
  - A score $s$ to determine whether neighboring words/phrase should be merged into a larger phrase, where $s = W^{\text{score}}p$
  - A new vector representation $p$ for the larger phrase

$$ p = f \left( W \begin{bmatrix} p_L \\ p_R \end{bmatrix} + b \right) $$

- Its class label. e.g., phrase types
- $W$ is recursively used everywhere in the tree
- Other models can be obtained by augmenting the recursive composition functions
Experiment I: Sentiment Analysis

<table>
<thead>
<tr>
<th>Model</th>
<th>Error rate (Positive/Negative)</th>
<th>Error rate (Fine-grained)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes (Socher et al., 2013b)</td>
<td>18.2%</td>
<td>59.0%</td>
</tr>
<tr>
<td>SVMs (Socher et al., 2013b)</td>
<td>20.6%</td>
<td>59.3%</td>
</tr>
<tr>
<td>Bigram Naïve Bayes (Socher et al., 2013b)</td>
<td>16.9%</td>
<td>58.1%</td>
</tr>
<tr>
<td>Word Vector Averaging (Socher et al., 2013b)</td>
<td>19.9%</td>
<td>67.3%</td>
</tr>
<tr>
<td>Recursive Neural Network (Socher et al., 2013b)</td>
<td>17.6%</td>
<td>56.8%</td>
</tr>
<tr>
<td>Matrix Vector-RNN (Socher et al., 2013b)</td>
<td>17.1%</td>
<td>55.6%</td>
</tr>
<tr>
<td>Recursive Neural Tensor Network (Socher et al., 2013b)</td>
<td>14.6%</td>
<td>54.3%</td>
</tr>
<tr>
<td>Paragraph Vector</td>
<td><strong>12.2%</strong></td>
<td><strong>51.3%</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>BoW (bnc) (Maas et al., 2011)</td>
<td>12.20%</td>
</tr>
<tr>
<td>BoW (bDt’c) (Maas et al., 2011)</td>
<td>11.77%</td>
</tr>
<tr>
<td>LDA (Maas et al., 2011)</td>
<td>32.58%</td>
</tr>
<tr>
<td>Full+BoW (Maas et al., 2011)</td>
<td>11.67%</td>
</tr>
<tr>
<td>Full+Unlabeled+BoW (Maas et al., 2011)</td>
<td>11.11%</td>
</tr>
<tr>
<td>WRRBM (Dahl et al., 2012)</td>
<td>12.58%</td>
</tr>
<tr>
<td>WRRBM + BoW (bnc) (Dahl et al., 2012)</td>
<td>10.77%</td>
</tr>
<tr>
<td>MNB-uni (Wang &amp; Manning, 2012)</td>
<td>16.45%</td>
</tr>
<tr>
<td>MNB-bi (Wang &amp; Manning, 2012)</td>
<td>13.41%</td>
</tr>
<tr>
<td>SVM-uni (Wang &amp; Manning, 2012)</td>
<td>13.05%</td>
</tr>
<tr>
<td>SVM-bi (Wang &amp; Manning, 2012)</td>
<td>10.84%</td>
</tr>
<tr>
<td>NBSVM-uni (Wang &amp; Manning, 2012)</td>
<td>11.71%</td>
</tr>
<tr>
<td>NBSVM-bi (Wang &amp; Manning, 2012)</td>
<td>8.78%</td>
</tr>
<tr>
<td>Paragraph Vector</td>
<td><strong>7.42%</strong></td>
</tr>
</tbody>
</table>

**Figure:** Stanford Sentiment Treebank dataset.

**Figure:** IMDB dataset.
Experiment II: Information Retrieval

Dataset

- 1,000,000 "triplets"
- Two paragraphs are results of the same query, whereas the third paragraph from a different query.

Performance

<table>
<thead>
<tr>
<th>Model</th>
<th>Error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vector Averaging</td>
<td>10.25%</td>
</tr>
<tr>
<td>Bag-of-words</td>
<td>8.10%</td>
</tr>
<tr>
<td>Bag-of-bigrams</td>
<td>7.28%</td>
</tr>
<tr>
<td>Weighted Bag-of-bigrams</td>
<td>5.67%</td>
</tr>
<tr>
<td>Paragraph Vector</td>
<td>3.82%</td>
</tr>
</tbody>
</table>
References

Le, Quoc V., and Tomas Mikolov.
Distributed representations of sentences and documents
*ICML 2014*

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean
Distributed representations of words and phrases and their compositionality
*NIPS 2013*

Bengio, Yoshua, Rejean Ducharme, Pascal Vincent, and Christian Janvin.
A neural probabilistic language model.
*The Journal of Machine Learning Research, 2003*

Richard Socher
Recursive Deep Learning for Natural Language Processing and Computer Vision
*PhD Thesis, Computer Science Department, Stanford University, 2014*