NLP and Word Embeddings

Word representation

deeplearning.ai
Word representation

\[ V = [a, aaron, \ldots, zulu, <UNK>] \]

1-hot representation

\[
\begin{align*}
\text{Man} & : \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ 0 \end{bmatrix} \\
\text{Woman} & : \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ 0 \end{bmatrix} \\
\text{King} & : \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ 0 \end{bmatrix} \\
\text{Queen} & : \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ 0 \end{bmatrix} \\
\text{Apple} & : \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ 0 \end{bmatrix} \\
\text{Orange} & : \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ 0 \end{bmatrix}
\end{align*}
\]

\[ |V| = 10,000 \]

I want a glass of orange ______.
I want a glass of apple______.
# Featurized representation: word embedding

<table>
<thead>
<tr>
<th>Gender</th>
<th>Royal</th>
<th>Age</th>
<th>Food</th>
<th>Size</th>
<th>Cost</th>
<th>Verb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Man</td>
<td>-1</td>
<td>0.01</td>
<td>0.03</td>
<td>0.04</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>Woman</td>
<td>1</td>
<td>0.02</td>
<td>0.02</td>
<td>0.01</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Man</th>
<th>Woman</th>
<th>King</th>
<th>Queen</th>
<th>Apple</th>
<th>Orange</th>
</tr>
</thead>
<tbody>
<tr>
<td>(5391)</td>
<td>(9853)</td>
<td>(4914)</td>
<td>(7157)</td>
<td>(456)</td>
<td>(6257)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Man</th>
<th>Woman</th>
<th>King</th>
<th>Queen</th>
<th>Apple</th>
<th>Orange</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.95</td>
<td>0.97</td>
<td>0.00</td>
<td>0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.93</td>
<td>0.95</td>
<td>-0.01</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.7</td>
<td>0.69</td>
<td>0.03</td>
<td>-0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.02</td>
<td>0.01</td>
<td>0.95</td>
<td>0.97</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

I want a glass of orange juice.
I want a glass of apple juice.

Andrew Ng
Visualizing word embeddings

[van der Maaten and Hinton, 2008. Visualizing data using t-SNE]
NLP and Word Embeddings

Using word embeddings
Named entity recognition example

Sally Johnson is an orange farmer

Robert Lin is an apple farmer

$\rightarrow$ 1B words $\rightarrow$ 100B words
$\rightarrow$ 100K words
Transfer learning and word embeddings

1. Learn word embeddings from large text corpus. (1-100B words)
   (Or download pre-trained embedding online.)

2. Transfer embedding to new task with smaller training set.
   (say, 100k words)

3. Optional: Continue to finetune the word embeddings with new data.
Relation to face encoding

\[ f(x^{(i)}) \]

\[ f(x^{(j)}) \]

[10,000]

\[ Q_1, \ldots, Q_{10,000} \]

\[ \hat{y} \]

[Taigman et. al., 2014. DeepFace: Closing the gap to human level performance]
## Analogies

<table>
<thead>
<tr>
<th></th>
<th>Man (5391)</th>
<th>Woman (9853)</th>
<th>King (4914)</th>
<th>Queen (7157)</th>
<th>Apple (456)</th>
<th>Orange (6257)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>-1</td>
<td>1</td>
<td>-0.95</td>
<td>0.97</td>
<td>0.00</td>
<td>0.01</td>
</tr>
<tr>
<td>Royal</td>
<td>0.01</td>
<td>0.02</td>
<td>0.93</td>
<td>0.95</td>
<td>-0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>Age</td>
<td>0.03</td>
<td>0.02</td>
<td>0.70</td>
<td>0.69</td>
<td>0.03</td>
<td>-0.02</td>
</tr>
<tr>
<td>Food</td>
<td>0.09</td>
<td>0.01</td>
<td>0.02</td>
<td>0.01</td>
<td>0.95</td>
<td>0.97</td>
</tr>
</tbody>
</table>

\[
\begin{bmatrix}
0.05 \\
-0.62 \\
0.14 \\
0.00
\end{bmatrix}
\approx
\begin{bmatrix}
-1 \\
1 \\
-0.95 \\
0.97
\end{bmatrix}
\]

\[
\begin{bmatrix}
0.01 \\
0.02 \\
0.93 \\
0.95
\end{bmatrix}
\approx
\begin{bmatrix}
0.01 \\
0.02 \\
0.93 \\
0.95
\end{bmatrix}
\]

\[
\begin{bmatrix}
0.03 \\
0.02 \\
0.70 \\
0.69
\end{bmatrix}
\approx
\begin{bmatrix}
0.03 \\
0.02 \\
0.70 \\
0.69
\end{bmatrix}
\]

\[
\begin{bmatrix}
0.09 \\
0.01 \\
0.02 \\
0.01
\end{bmatrix}
\approx
\begin{bmatrix}
0.09 \\
0.01 \\
0.02 \\
0.01
\end{bmatrix}
\]

[Mikolov et al., 2013, Linguistic regularities in continuous space word representations]
Analogies using word vectors

- Formula: $e_{\text{man}} - e_{\text{woman}} \approx e_{\text{king}} - e_{\text{queen}}$
- Example: $3000 \rightarrow 2D$
- Algorithm: Find word $w$ with arg max $\text{sim}(e_w, e_{\text{king}} - e_{\text{man}} + e_{\text{woman}})$
Cosine similarity

\[ \text{sim}(e_w, e_{\text{king}} - e_{\text{man}} + e_{\text{woman}}) \]

\[ \text{Sim}(u, v) = \frac{u^T v}{\|u\| \|v\|} \]

Man:Woman as Boy:Girl
Ottawa:Canada as Nairobi:Kenya
Big:Bigger as Tall:Taller
Yen:Japan as Ruble:Russia
NLP and Word Embeddings

Embedding matrix
Embedding matrix

In practice, use specialized function to look up an embedding.

Andrew Ng
NLP and Word Embeddings

Learning word embeddings
Neural language model

I want a glass of orange.  

[Neural network diagram showing the process of predicting words in a sentence]

[Bengio et. al., 2003, A neural probabilistic language model]
I want a glass of orange juice to go along with my cereal.

Context: Last 4 words.

- 4 words on left & right
- Last 1 word
- Nearby 1 word

skip gram
NLP and Word Embeddings

Word2Vec
Skip-grams

I want a glass of orange juice to go along with my cereal.

[Mikolov et. al., 2013. Efficient estimation of word representations in vector space.]
Model

Vocab size = 10,000k

\[ p(t|c) = \frac{e^{O_t \cdot E_c}}{\sum_{j=1}^{10000} e^{O_j \cdot E_c}} \]

\[ d(\hat{y}, y) = - \sum_{i=1}^{4834} y_i \log \hat{y_i} \]

\[ y = \begin{bmatrix} \vdots \\ 4834 \end{bmatrix} \]
Problems with softmax classification

\[ p(t|c) = \frac{e^{\theta^T tc}}{\sum_{j=1}^{10,000} e^{\theta^T j c}} \]

How to sample the context \( c \)?

\( \Rightarrow \) the, of, a, and, to, ...

\( \Rightarrow \) orange, apple, durian

\[ P(c) \]

Andrew Ng
NLP and Word Embeddings

Negative sampling
Defining a new learning problem

I want a glass of orange juice to go along with my cereal.

[Mikolov et al., 2013. Distributed representation of words and phrases and their compositionality]
Model

Softmax: \[ p(t|c) = \frac{e^{\theta_t^T e_c}}{\sum_{j=1}^{10,000} e^{\theta_j^T e_c}} \]

\[ P(y=1 | c, t) = \sigma(\Theta_e^T e_c) \]
Selecting negative examples

<table>
<thead>
<tr>
<th>context</th>
<th>word</th>
<th>target?</th>
</tr>
</thead>
<tbody>
<tr>
<td>orange</td>
<td>juice</td>
<td>1</td>
</tr>
<tr>
<td>orange</td>
<td>king</td>
<td>0</td>
</tr>
<tr>
<td>orange</td>
<td>book</td>
<td>0</td>
</tr>
<tr>
<td>orange</td>
<td>the</td>
<td>0</td>
</tr>
<tr>
<td>orange</td>
<td>of</td>
<td>0</td>
</tr>
</tbody>
</table>

\[
P(\omega_i) = \frac{f(\omega_i)^{3/4}}{\sum_{j=1}^{10,000} f(\omega_j)^{3/4}} \frac{1}{|V|}
\]
NLP and Word Embeddings

GloVe word vectors
GloVe (global vectors for word representation)

I want a glass of orange juice to go along with my cereal.

\[ X_{ij} = \# \text{ times } i \text{ appears in context of } j. \]

\[ X_{ij} = X_{ji} \]
Model

Minimize

\[ \sum_{i=1}^{10,000} \sum_{j=1}^{10,000} f(x_{ij})(\Theta_i^T e_j + b_i + b_j' - \log(x_{ij}))^2 \]

Weight term

\[ f(x_{ij}) = 0 \text{ or } x_{ij} = 0. \]

"\( \Theta_i^T e_j \)" is symmetric

\[ 0_i \text{, } e_j \text{ are symmetric} \]

\[ E_{f_{\text{final}}} = E_{W_{f_{\text{final}}}} + \Theta \frac{1}{2} \]

Andrew Ng
A note on the featurization view of word embeddings

minimize $\sum_{i=1}^{10,000} \sum_{j=1}^{10,000} f(X_{ij}) (\theta_i^T e_j + b_i - b'_j - \log X_{ij})^2$

<table>
<thead>
<tr>
<th></th>
<th>Man</th>
<th>Woman</th>
<th>King</th>
<th>Queen</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>-1</td>
<td>1</td>
<td>-0.95</td>
<td>0.97</td>
</tr>
<tr>
<td>Royal</td>
<td>0.01</td>
<td>0.02</td>
<td>0.93</td>
<td>0.95</td>
</tr>
<tr>
<td>Age</td>
<td>0.03</td>
<td>0.02</td>
<td>0.70</td>
<td>0.69</td>
</tr>
<tr>
<td>Food</td>
<td>0.09</td>
<td>0.01</td>
<td>0.02</td>
<td>0.01</td>
</tr>
</tbody>
</table>
Sentiment classification problem

\[ x \rightarrow y \]

The dessert is excellent.  

Service was quite slow.  

Good for a quick meal, but nothing special.  

Completely lacking in good taste, good service, and good ambience.

10,000 \Rightarrow 100,000 words
Simple sentiment classification model

The dessert is excellent

8928 2468 4694 3180

“The dessert is excellent.

“Completely lacking in good taste, good service, and good ambience.”

Andrew Ng
RNN for sentiment classification

$\hat{y}$

softmax

$e_{330}$
$E$

$a^{<10>}$

$e_{3882}$
$E$

$\ldots$

$e_{4427}$
$E$

$\ldots$

$a^{<4>}$

$e_{4966}$
$E$

$\ldots$

$a^{<2>}$

$e_{1852}$
$E$

$\ldots$

$a^{<0>}$

Completely lacking in good ambience,

many-to-one "not good"

Andrew Ng
NLP and Word Embeddings

Debiasing word embeddings
The problem of bias in word embeddings

Man:Woman as King:Queen

Man:Computer_Programmer as Woman:Homemaker

Father:Doctor as Mother:Nurse

Word embeddings can reflect gender, ethnicity, age, sexual orientation, and other biases of the text used to train the model.

[Bolukbasi et. al., 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings]
Addressing bias in word embeddings

1. Identify bias direction.

2. Neutralize: For every word that is not definitional, project to get rid of bias.

3. Equalize pairs.

[Bolukbasi et. al., 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings]