Word2vec and beyond

presented by Eleni Triantafillou

March 1, 2016
The Big Picture

There is a long history of word representations

- Techniques from information retrieval: Latent Semantic Analysis (LSA)
- Self-Organizing Maps (SOM)
- Distributional count-based methods
- Neural Language Models

Important take-aways:

1. Don’t need deep models to get good embeddings
2. Count-based models and neural net predictive models are not qualitatively different

source:
http://gavagai.se/blog/2015/09/30/a-brief-history-of-word-embeddings/
Continuous Word Representations

- Contrast with simple n-gram models (words as atomic units)
- Simple models have the potential to perform very well...
- ... if we had enough data
- Need more complicated models
- Continuous representations take better advantage of data by modelling the similarity between the words
Continuous Representations

source: http://www.codeproject.com/Tips/788739/Visualization-of-High-Dimensional-Data-using-t-SNE
Learn to predict surrounding words

Use a large training corpus to maximize:

$$\frac{1}{T} \sum_{t=1}^{T} \sum_{-c \leq j \leq c, \ j \neq 0} \log p(w_{t+j} | w_t)$$

where:

- $T$: training set size
- $c$: context size
- $w_j$: vector representation of the $j_{th}$ word
Skip Gram: Think of it as a Neural Network

Learn $W$ and $W'$ in order to maximize previous objective.

source: "word2vec parameter learning explained." ([4])
CBOW

source: "word2vec parameter learning explained." ([4])
word2vec Experiments

- Evaluate how well syntactic/semantic word relationships are captured
- Understand effect of increasing training size / dimensionality
- Microsoft Research Sentence Completion Challenge
Table 1: *Examples of five types of semantic and nine types of syntactic questions in the Semantic-Syntactic Word Relationship test set.*

<table>
<thead>
<tr>
<th>Type of relationship</th>
<th>Word Pair 1</th>
<th>Word Pair 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common capital city</td>
<td>Athens</td>
<td>Greece</td>
</tr>
<tr>
<td>All capital cities</td>
<td>Astana</td>
<td>Kazakhstan</td>
</tr>
<tr>
<td>Currency</td>
<td>Angola</td>
<td>kwanza</td>
</tr>
<tr>
<td>City-in-state</td>
<td>Chicago</td>
<td>Illinois</td>
</tr>
<tr>
<td>Man-Woman</td>
<td>brother</td>
<td>sister</td>
</tr>
<tr>
<td>Adjective to adverb</td>
<td>apparent</td>
<td>apparently</td>
</tr>
<tr>
<td>Opposite</td>
<td>possibly</td>
<td>impossibly</td>
</tr>
<tr>
<td>Comparative</td>
<td>great</td>
<td>greater</td>
</tr>
<tr>
<td>Superlative</td>
<td>easy</td>
<td>easiest</td>
</tr>
<tr>
<td>Present Participle</td>
<td>think</td>
<td>thinking</td>
</tr>
<tr>
<td>Nationality adjective</td>
<td>Switzerland</td>
<td>Swiss</td>
</tr>
<tr>
<td>Past tense</td>
<td>walking</td>
<td>walked</td>
</tr>
<tr>
<td>Plural nouns</td>
<td>mouse</td>
<td>mice</td>
</tr>
<tr>
<td>Plural verbs</td>
<td>work</td>
<td>works</td>
</tr>
<tr>
<td></td>
<td>Oslo</td>
<td>Stockton</td>
</tr>
<tr>
<td></td>
<td>Harare</td>
<td>grandson</td>
</tr>
<tr>
<td></td>
<td>Norway</td>
<td>Zimbabwe</td>
</tr>
<tr>
<td></td>
<td>rial</td>
<td>California</td>
</tr>
<tr>
<td></td>
<td>rapid</td>
<td>unethically</td>
</tr>
<tr>
<td></td>
<td>ethical</td>
<td>unethical</td>
</tr>
<tr>
<td></td>
<td>tough</td>
<td>tougher</td>
</tr>
<tr>
<td></td>
<td>lucky</td>
<td>luckiest</td>
</tr>
<tr>
<td></td>
<td>read</td>
<td>reading</td>
</tr>
<tr>
<td></td>
<td>Cambodia</td>
<td>Cambodian</td>
</tr>
<tr>
<td></td>
<td>swimming</td>
<td>swim</td>
</tr>
<tr>
<td></td>
<td>dollar</td>
<td>dollars</td>
</tr>
<tr>
<td></td>
<td>speak</td>
<td>speaks</td>
</tr>
</tbody>
</table>
Table 4: *Comparison of publicly available word vectors on the Semantic-Syntactic Word Relationship test set, and word vectors from our models. Full vocabularies are used.*

<table>
<thead>
<tr>
<th>Model</th>
<th>Vector Dimensionality</th>
<th>Training words</th>
<th>Accuracy [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Semantic</td>
</tr>
<tr>
<td>Collobert-Weston NNLM</td>
<td>50</td>
<td>660M</td>
<td>9.3</td>
</tr>
<tr>
<td>Turian NNLM</td>
<td>50</td>
<td>37M</td>
<td>1.4</td>
</tr>
<tr>
<td>Turian NNLM</td>
<td>200</td>
<td>37M</td>
<td>1.4</td>
</tr>
<tr>
<td>Mnih NNLM</td>
<td>50</td>
<td>37M</td>
<td>1.8</td>
</tr>
<tr>
<td>Mnih NNLM</td>
<td>100</td>
<td>37M</td>
<td>3.3</td>
</tr>
<tr>
<td>Mikolov RNNLM</td>
<td>80</td>
<td>320M</td>
<td>4.9</td>
</tr>
<tr>
<td>Mikolov RNNLM</td>
<td>640</td>
<td>320M</td>
<td>8.6</td>
</tr>
<tr>
<td>Huang NNLM</td>
<td>50</td>
<td>990M</td>
<td>13.3</td>
</tr>
<tr>
<td>Our NNLM</td>
<td>20</td>
<td>6B</td>
<td>12.9</td>
</tr>
<tr>
<td>Our NNLM</td>
<td>50</td>
<td>6B</td>
<td>27.9</td>
</tr>
<tr>
<td>Our NNLM</td>
<td>100</td>
<td>6B</td>
<td>34.2</td>
</tr>
<tr>
<td>CBOW</td>
<td>300</td>
<td>783M</td>
<td>15.5</td>
</tr>
<tr>
<td>Skip-gram</td>
<td>300</td>
<td>783M</td>
<td>50.0</td>
</tr>
</tbody>
</table>
Table 8: *Examples of the word pair relationships, using the best word vectors from Table 4 (Skip-gram model trained on 783M words with 300 dimensionality).*

<table>
<thead>
<tr>
<th>Relationship</th>
<th>Example 1</th>
<th>Example 2</th>
<th>Example 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>France - Paris</td>
<td>Italy: Rome</td>
<td>Japan: Tokyo</td>
<td>Florida: Tallahassee</td>
</tr>
<tr>
<td>big - bigger</td>
<td>small: larger</td>
<td>cold: colder</td>
<td>quick: quicker</td>
</tr>
<tr>
<td>Miami - Florida</td>
<td>Baltimore: Maryland</td>
<td>Dallas: Texas</td>
<td>Kona: Hawaii</td>
</tr>
<tr>
<td>Einstein - scientist</td>
<td>Messi: midfielder</td>
<td>Mozart: violinist</td>
<td>Picasso: painter</td>
</tr>
<tr>
<td>Sarkozy - France</td>
<td>Berlusconi: Italy</td>
<td>Merkel: Germany</td>
<td>Koizumi: Japan</td>
</tr>
<tr>
<td>copper - Cu</td>
<td>zinc: Zn</td>
<td>gold: Au</td>
<td>uranium: plutonium</td>
</tr>
<tr>
<td>Berlusconi - Silvio</td>
<td>Sarkozy: Nicolas</td>
<td>Putin: Medvedev</td>
<td>Obama: Barack</td>
</tr>
<tr>
<td>Microsoft - Windows</td>
<td>Google: Android</td>
<td>IBM: Linux</td>
<td>Apple: iPhone</td>
</tr>
<tr>
<td>Microsoft - Ballmer</td>
<td>Google: Yahoo</td>
<td>IBM: McNealy</td>
<td>Apple: Jobs</td>
</tr>
<tr>
<td>Japan - sushi</td>
<td>Germany: bratwurst</td>
<td>France: tapas</td>
<td>USA: pizza</td>
</tr>
</tbody>
</table>
Microsoft Research Sentence Completion

Table 7: Comparison and combination of models on the Microsoft Sentence Completion Challenge.

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Accuracy [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>4-gram [32]</td>
<td>39</td>
</tr>
<tr>
<td>Average LSA similarity [32]</td>
<td>49</td>
</tr>
<tr>
<td>Log-bilinear model [24]</td>
<td>54.8</td>
</tr>
<tr>
<td>RNNLMs [19]</td>
<td>55.4</td>
</tr>
<tr>
<td>Skip-gram</td>
<td>48.0</td>
</tr>
<tr>
<td>Skip-gram + RNNLMs</td>
<td><strong>58.9</strong></td>
</tr>
</tbody>
</table>
Linguistic Regularities

▶ "king" - "man" + "woman" = "queen"!
▶ Demo
▶ Check out gensim (python library for topic modelling): https://radimrehurek.com/gensim/models/word2vec.html
Multimodal Word Embeddings: Motivation

Are these two objects similar?
Multimodal Word Embeddings: Motivation

And these?
What do you think should be the case?

\[
sim(\text{pizza}, \text{clock}) \quad < \quad \sim(\text{pizza}, \text{banana})
\]

or

\[
sim(\text{pizza}, \text{clock}) \quad > \quad \sim(\text{pizza}, \text{banana})
\]
When do we need image features?

It’s surely task-specific. In many cases can benefit from visual features!

- Text-based Image Retrieval
- Visual Paraphrasing
- Common Sense Assertion Classification
- They are better-suited for zero shot learning (learn mapping between text and images)
Two Multimodal Word Embeddings approaches...

1. Combining Language and Vision with a Multimodal Skip-gram Model (Lazaridou et al, 2013)
Two Multimodal Word Embeddings approaches...

1. Combining Language and Vision with a Multimodal Skip-gram Model (Lazaridou et al, 2013)
Multimodal Skip-Gram

- **The main idea**: Use visual features for the (very) small subset of the training data for which images are available.
- Visual vectors are obtained by CNN and are fixed during training!
- Recall, Skip-Gram objective:

\[
L_{\text{ling}}(w_t) = \sum_{t=1}^{T} \sum_{-c \leq j \leq c, j \neq 0} \log(p(w_{t+j}|w_t))
\]

- New Multimodal Skip-Gram objective:

\[
L = \frac{1}{T} \sum_{t=1}^{T} (L_{\text{ling}}(w_t) + L_{\text{vision}}(w_t)),
\]

where

- \(L_{\text{vision}}(w_t) = 0\) if \(w_t\) does not have an entry in ImageNet, and otherwise

\[
L_{\text{vision}}(w_t) = - \sum_{w' \sim P(w)} \max(0, \gamma - \cos(u_{w_t}, v_{w_t}) + \cos(u_{w_t}, v_{w'}))
\]
# Multimodal Skip-Gram: An example

## Training Set

<table>
<thead>
<tr>
<th>Words</th>
<th>Image Available?</th>
</tr>
</thead>
<tbody>
<tr>
<td>pizza</td>
<td>yes</td>
</tr>
<tr>
<td>cat</td>
<td>yes</td>
</tr>
<tr>
<td>clock</td>
<td>yes</td>
</tr>
<tr>
<td>love</td>
<td>no</td>
</tr>
<tr>
<td>oven</td>
<td>no</td>
</tr>
</tbody>
</table>
Multimodal Skip-Gram: An example

Embeddings for words (init)

“pizza”

“love”

“cat”

“oven”

“clock”
Multimodal Skip-Gram: An example

Embeddings for words (training)

- "pizza"
- "love"
- "cat"
- "oven"
- "clock"
Multimodal Skip-Gram: An example

Embeddings for words (trained)

“pizza”

“oven”

“love”

“clock”

“cat”
Multimodal Skip-Gram: An example

Multi-modal Embeddings

“pizza”
“love”
“oven”
“clock”
“cat”
Multimodal Skip-Gram: An example
Multimodal Skip-Gram: An example

Multi-modal Embeddings

- "clock"
- "pizza"
- "oven"
- "love"
- "cat"
Multimodal Skip-Gram: An example

Nearest Neighbors of ‘pizza’

‘pizza’ and ‘clock’ are now neighbors!

“clock”

“pizza”

“oven”

“love”

“cat”
Multimodal Skip-Gram: Comparing to Human Judgements

<table>
<thead>
<tr>
<th>Model</th>
<th>MEN 100%</th>
<th>MEN 42%</th>
<th>Simlex-999 100%</th>
<th>Simlex-999 29%</th>
<th>SemSim 100%</th>
<th>SemSim 85%</th>
<th>VisSim 100%</th>
<th>VisSim 85%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kiela and Bottou</td>
<td>-</td>
<td>0.74</td>
<td>-</td>
<td>0.33</td>
<td>-</td>
<td>0.60</td>
<td>-</td>
<td>0.50</td>
</tr>
<tr>
<td>Bruni et al.</td>
<td>-</td>
<td>0.77</td>
<td>-</td>
<td>0.44</td>
<td>-</td>
<td>0.69</td>
<td>-</td>
<td>0.56</td>
</tr>
<tr>
<td>Silberer and Lapata</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.70</td>
<td>-</td>
<td>0.64</td>
<td>-</td>
</tr>
<tr>
<td>CNN features</td>
<td>-</td>
<td>0.62</td>
<td>-</td>
<td>0.54</td>
<td>-</td>
<td>0.55</td>
<td>-</td>
<td>0.56</td>
</tr>
<tr>
<td>Skip-gram</td>
<td>0.70</td>
<td>0.68</td>
<td>0.33</td>
<td>0.29</td>
<td>0.62</td>
<td>0.62</td>
<td>0.48</td>
<td>0.48</td>
</tr>
<tr>
<td>Concatenation</td>
<td>-</td>
<td>0.74</td>
<td>-</td>
<td>0.46</td>
<td>-</td>
<td>0.68</td>
<td>-</td>
<td>0.60</td>
</tr>
<tr>
<td>SVD</td>
<td>0.61</td>
<td>0.74</td>
<td>0.28</td>
<td>0.46</td>
<td>0.65</td>
<td>0.68</td>
<td>0.58</td>
<td>0.60</td>
</tr>
<tr>
<td>MMSkip-gram-A</td>
<td>0.75</td>
<td>0.74</td>
<td>0.37</td>
<td>0.50</td>
<td>0.72</td>
<td>0.72</td>
<td>0.63</td>
<td>0.63</td>
</tr>
<tr>
<td>MMSkip-gram-B</td>
<td>0.74</td>
<td>0.76</td>
<td>0.40</td>
<td>0.53</td>
<td>0.66</td>
<td>0.68</td>
<td>0.60</td>
<td>0.60</td>
</tr>
</tbody>
</table>

**MEN**: general relatedness (”pickles”, ”hamburgers”), **Simplex-999**: taxonomic similarity (”pickles”, ”food”), **SemSim**: Semantic similarity (”pickles”, ”onions”), **VisSim**: Visual Similarity (”pen”, ”screwdriver”)
Multimodal Skip-Gram: Examples of Nearest Neighbors

Only “donut” and “owl” trained with direct visual information.

<table>
<thead>
<tr>
<th>Target</th>
<th>SKIP-GRAM</th>
<th>MMSkip-gram-A</th>
<th>MMSkip-gram-B</th>
</tr>
</thead>
<tbody>
<tr>
<td>donut</td>
<td>fridge, diner, candy</td>
<td>pizza, sushi, sandwich</td>
<td>pizza, sushi, sandwich</td>
</tr>
<tr>
<td>owl</td>
<td>pheasant, woodpecker, squirrel</td>
<td>eagle, woodpecker, falcon</td>
<td>eagle, falcon, hawk</td>
</tr>
<tr>
<td>mural</td>
<td>sculpture, painting, portrait</td>
<td>painting, portrait, sculpture</td>
<td>painting, portrait, sculpture</td>
</tr>
<tr>
<td>tobacco</td>
<td>coffee, cigarette, corn</td>
<td>cigarette, cigar, corn</td>
<td>cigarette, cigar, smoking</td>
</tr>
<tr>
<td>depth</td>
<td>size, bottom, meter</td>
<td>sea, underwater, level</td>
<td>sea, size, underwater</td>
</tr>
<tr>
<td>chaos</td>
<td>anarchy, despair, demon</td>
<td>demon, anarchy, destruction</td>
<td>demon, anarchy, shadow</td>
</tr>
</tbody>
</table>
Multimodal Skip-Gram: Zero-shot image labelling and image retrieval

<table>
<thead>
<tr>
<th></th>
<th>P@1</th>
<th>P@2</th>
<th>P@10</th>
<th>P@20</th>
<th>P@50</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SKIP-GRAM</strong></td>
<td>1.5</td>
<td>2.6</td>
<td>14.2</td>
<td>23.5</td>
<td>36.1</td>
</tr>
<tr>
<td><strong>MMSKIP-GRAM-A</strong></td>
<td>2.1</td>
<td>3.7</td>
<td>16.7</td>
<td>24.6</td>
<td>37.6</td>
</tr>
<tr>
<td><strong>MMSKIP-GRAM-B</strong></td>
<td>2.2</td>
<td>5.1</td>
<td>20.2</td>
<td>28.5</td>
<td>43.5</td>
</tr>
</tbody>
</table>

Table 3: Percentage precision@k results in the zero-shot image labeling task.

<table>
<thead>
<tr>
<th></th>
<th>P@1</th>
<th>P@2</th>
<th>P@10</th>
<th>P@20</th>
<th>P@50</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SKIP-GRAM</strong></td>
<td>1.9</td>
<td>3.3</td>
<td>11.5</td>
<td>18.5</td>
<td>30.4</td>
</tr>
<tr>
<td><strong>MMSKIP-GRAM-A</strong></td>
<td>1.9</td>
<td>3.2</td>
<td>13.9</td>
<td>20.2</td>
<td>33.6</td>
</tr>
<tr>
<td><strong>MMSKIP-GRAM-B</strong></td>
<td>1.9</td>
<td>3.8</td>
<td>13.2</td>
<td>22.5</td>
<td>38.3</td>
</tr>
</tbody>
</table>

Table 4: Percentage precision@k results in the zero-shot image retrieval task.
Multimodal Skip-Gram: Survey to evaluate on Abstract Words

**Metric**: Proportion (percentage) of words for which number votes in favour of “neighbour” image significantly above chance.

**Unseen**: Discard words for which visual info was accessible during training.

<table>
<thead>
<tr>
<th></th>
<th>global</th>
<th>words</th>
<th>unseen</th>
<th>words</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>48%</td>
<td>198</td>
<td>30%</td>
<td>127</td>
</tr>
<tr>
<td>concrete</td>
<td>73%</td>
<td>99</td>
<td>53%</td>
<td>30</td>
</tr>
<tr>
<td>abstract</td>
<td>23%</td>
<td>99</td>
<td>23%</td>
<td>97</td>
</tr>
</tbody>
</table>
Multimodal Skip-Gram: Survey to evaluate on Abstract Words

Left: subject preferred the nearest neighbour to the random image

- freedom
- theory
- wrong

- god
- together
- place
Two Multimodal Word Embeddings approaches...

1. Combining Language and Vision with a Multimodal Skip-gram Model (Lazaridou et al, 2013)
Visual Word2Vec (vis-w2v): Motivation

w2v : farther
- eating
- stares at

vis-w2v : closer
- eating
- stares at

Word Embedding

girl eating ice cream

girl stares at ice cream
Visual Word2Vec (vis-w2v): Approach

- Multimodal train set: tuples of (description, abstract scene)
- **Finetune** word2vec to add visual features obtained by abstract scenes (clipart)
- Obtain surrogate (visual) classes by clustering those features
- \( W_I \): initialized from word2vec
- \( N_K \): number of clusters of abstract scene features

![Diagram of Visual Word2Vec (vis-w2v) model]

\( O \) \( \cdots \) \( O \) \( N_K \)

\( W_O \)

\( H \) \( \cdots \) \( O \) \( N_H \)

\( W_I \)

\( w_{i1} \) \( \cdots \) \( O \) \( N_V \) \( \cdots \) \( w_{i1} \) \( \cdots \) \( O \) \( N_V \)

\( \text{Output} \)

\( \text{Hidden} \)

\( \text{Input} \)
Clustering abstract scenes

Interestingly, “prepare to cut”, “hold”, “give” are clustered together with “stare at” etc. It would be hard to infer these semantic relationships from text alone.
Visual Word2Vec (vis-w2v): Relationship to CBOW (word2vec)

Surrogate labels play the role of visual context.
Figure 5: The visual paraphrasing task is to identify if two textual descriptions are paraphrases of each other. Shown above are three positive instances, *i.e.*, the descriptions (left, right) actually talk about the same scene (center). Green boxes show two cases where vis-w2v correctly predicts and w2v does not, while red box shows the case where both vis-w2v and w2v predict incorrectly. Note that the red instance is tough as the textual descriptions do not intuitively seem to be talking about the same scene, even for a human reader.
Visual Word2Vec (vis-w2v): Visual Paraphrasing Results

<table>
<thead>
<tr>
<th>Approach</th>
<th>Visual Paraphrasing AP (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>w2v-wiki</td>
<td>94.1</td>
</tr>
<tr>
<td>w2v-wiki</td>
<td>94.4</td>
</tr>
<tr>
<td>w2v-coco</td>
<td>94.6</td>
</tr>
<tr>
<td>vis-w2v-wiki</td>
<td>95.1</td>
</tr>
<tr>
<td>vis-w2v-coco</td>
<td><strong>95.3</strong></td>
</tr>
</tbody>
</table>

**Table:** Performance on visual paraphrasing task
Visual Word2Vec (vis-w2v): Common Sense Assertion Classification Results

Given a tuple (Primary Object, Relation, Secondary Object), decide if it is plausible or not.

<table>
<thead>
<tr>
<th>Approach</th>
<th>common sense AP (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>w2v-coco</td>
<td>72.2</td>
</tr>
<tr>
<td>w2v-wiki</td>
<td>68.1</td>
</tr>
<tr>
<td>w2v-coco + vision</td>
<td>73.6</td>
</tr>
<tr>
<td>vis-w2v-coco (shared)</td>
<td>74.5</td>
</tr>
<tr>
<td>vis-w2v-coco (shared) + vision</td>
<td>74.2</td>
</tr>
<tr>
<td>vis-w2v-coco (separate)</td>
<td>74.8</td>
</tr>
<tr>
<td>vis-w2v-coco (separate) + vision</td>
<td>75.2</td>
</tr>
<tr>
<td>vis-w2v-wiki (shared)</td>
<td>72.2</td>
</tr>
<tr>
<td>vis-w2v-wiki (separate)</td>
<td>74.2</td>
</tr>
</tbody>
</table>

Table: Performance on the common sense task
Thank you!

\[
\begin{bmatrix}
-0.0665592 & -0.0431451 & \ldots & -0.05182673 & -0.07418852 & -0.04472357 \\
0.02315103 & -0.04419742 & -0.01104935 \\
\end{bmatrix}
\]

\[
\begin{bmatrix}
0.08773034 & 0.00566679 & \ldots & 0.03735885 & -0.04323553 & 0.02130294 \\
-0.09108844 & -0.05708769 & 0.04659363 \\
\end{bmatrix}
\]
Bibliography


