Insight
Centre for Data Analytics

#InsightDL2018

DEEP LEARNING WORKSHOP

Dublin City University
21-22 May 2018

Day 1 Lecture 5

Word Embeddings

Xavier Giro-i-Nieto
xavier.giro@upc.edu

Associate Professor
Intelligent Data Science and Artificial Intelligence Center
Universitat Politecnica de Catalunya (UPC)

DCU

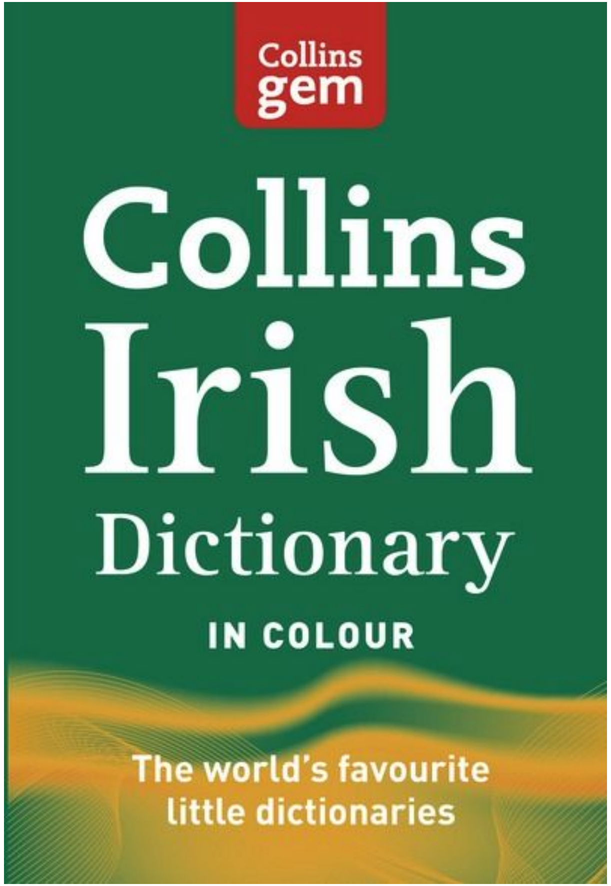
Acknowledgements

Antonio Bonafonte

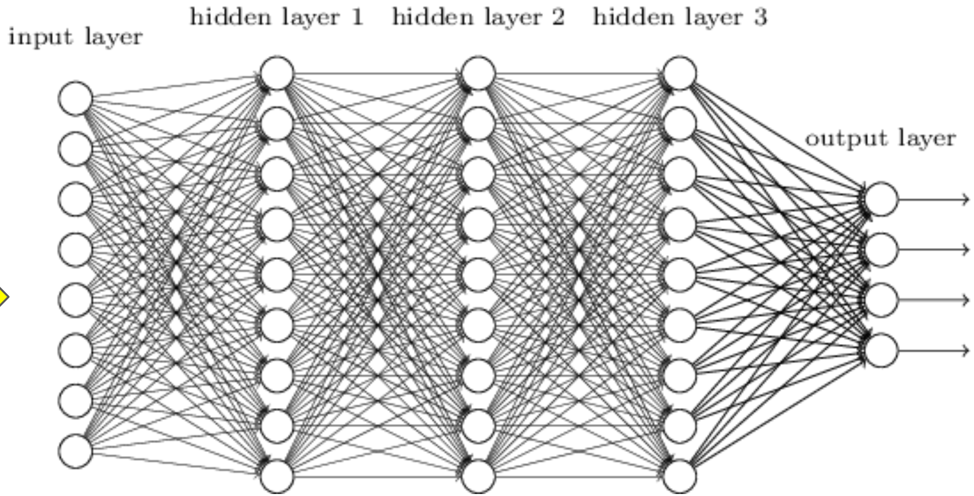
Santiago Pascual

TALP
UPC
TALP Research Center

Motivation


Collins gem
Collins Irish Dictionary
IN COLOUR
The world's favourite little dictionaries

?


input layer
hidden layer 1
hidden layer 2
hidden layer 3
output layer

One-hot encoding

Example: letters. |V| = 30

```
'a': x = 1
'b': x = 2
'c': x = 3
.
.
.
'.': x = 30
```

One-hot encoding

Example: letters. |V| = 30

```
'a': x = 1
'b': x = 2
'c': x = 3
.
.
.

'.': x = 30
```

We impose fake range ordering

One-hot encoding

Example: letters. |V| = 30

```
'a': x^T = [1,0,0, ..., 0]
'b': x^T = [0,1,0, ..., 0]
'c': x^T = [0,0,1, ..., 0]
.
.
.
'.': x^T = [0,0,0, ..., 1]
```

One-hot encoding

Example: words.

cat: x^T = [1,0,0, ..., 0]
dog: x^T = [0,1,0, ..., 0]
.
.
house: x^T = [0,0,0, …,0,1,0,...,0]
.
.
.

Number of words, |V| ?

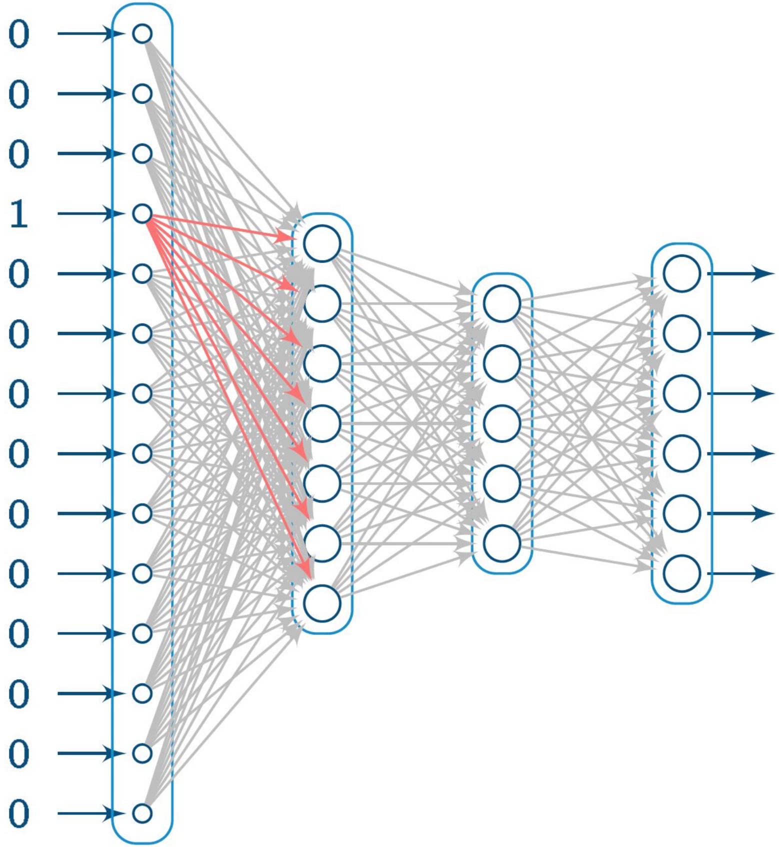
B2: 5K
C2: 18K
LVSR: 50-100K
Wikipedia (1.6B): 400K
Crawl data (42B): 2M

One-hot encoding

- Large dimensionality
- Sparse representation (mostly zeros)
- Blind representation
    - Only operators: ‘!=’ and ‘==’

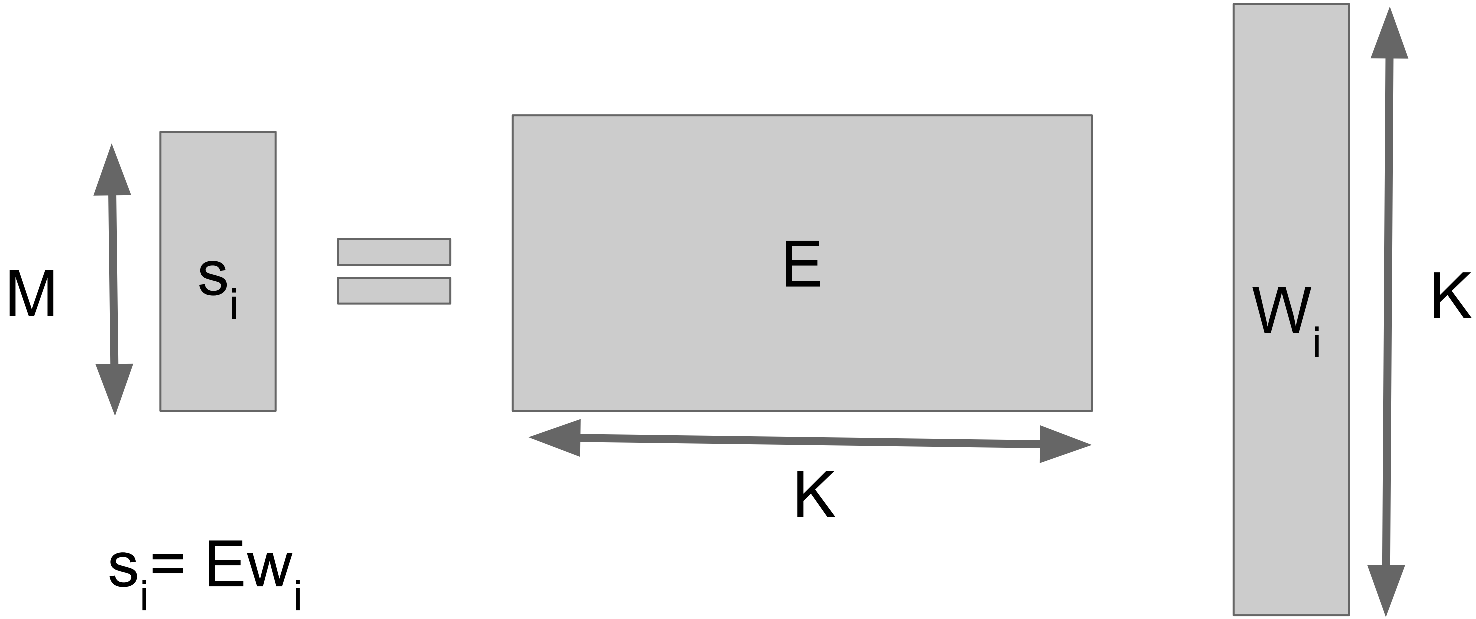
Projection to a word embedding

The one-hot is linearly projected to a embedded space of lower dimension with matrix E for learned weights (=fully connected).



Projection to a word embedding

The one-hot is linearly projected to a embedded space of lower dimension with matrix E for learned weights (=fully connected).



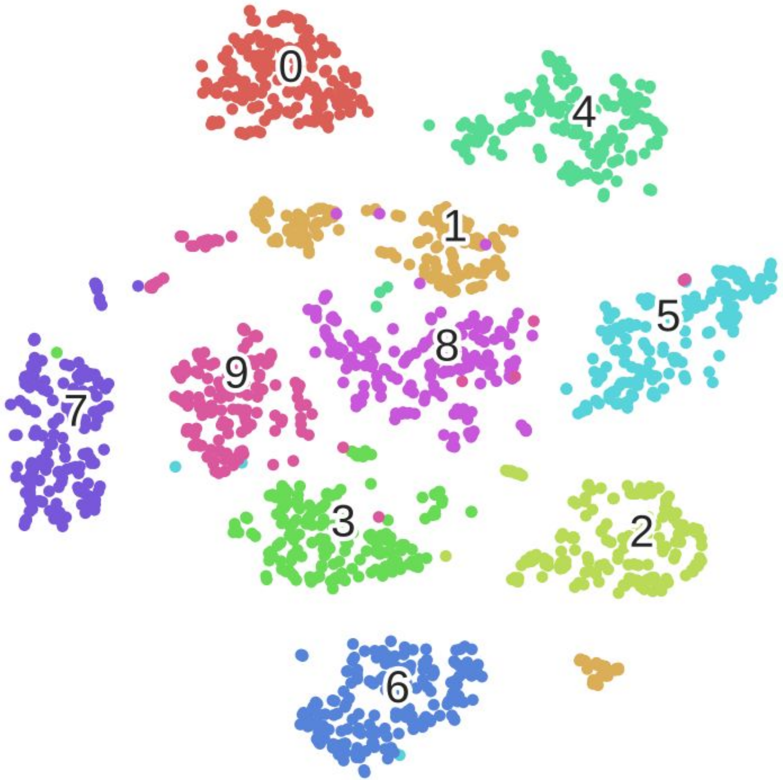
$s_i = Ew_i$

Projection to a word embedding

t-SNE

Embed high dimensional data points (i.e. feature codes) so that pairwise distances are preserved in local neighborhoods.

s_i



Maaten & Hinton. [Visualizing High-Dimensional Data using t-SNE](). Journal of Machine Learning Research (2008).

Projection to a word embedding

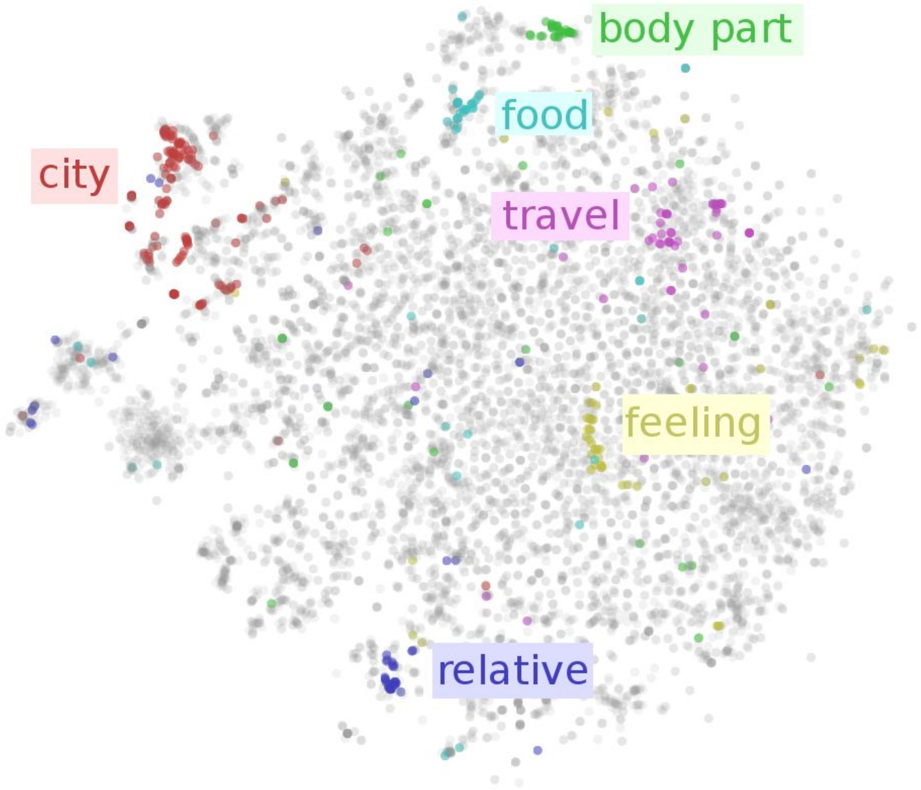
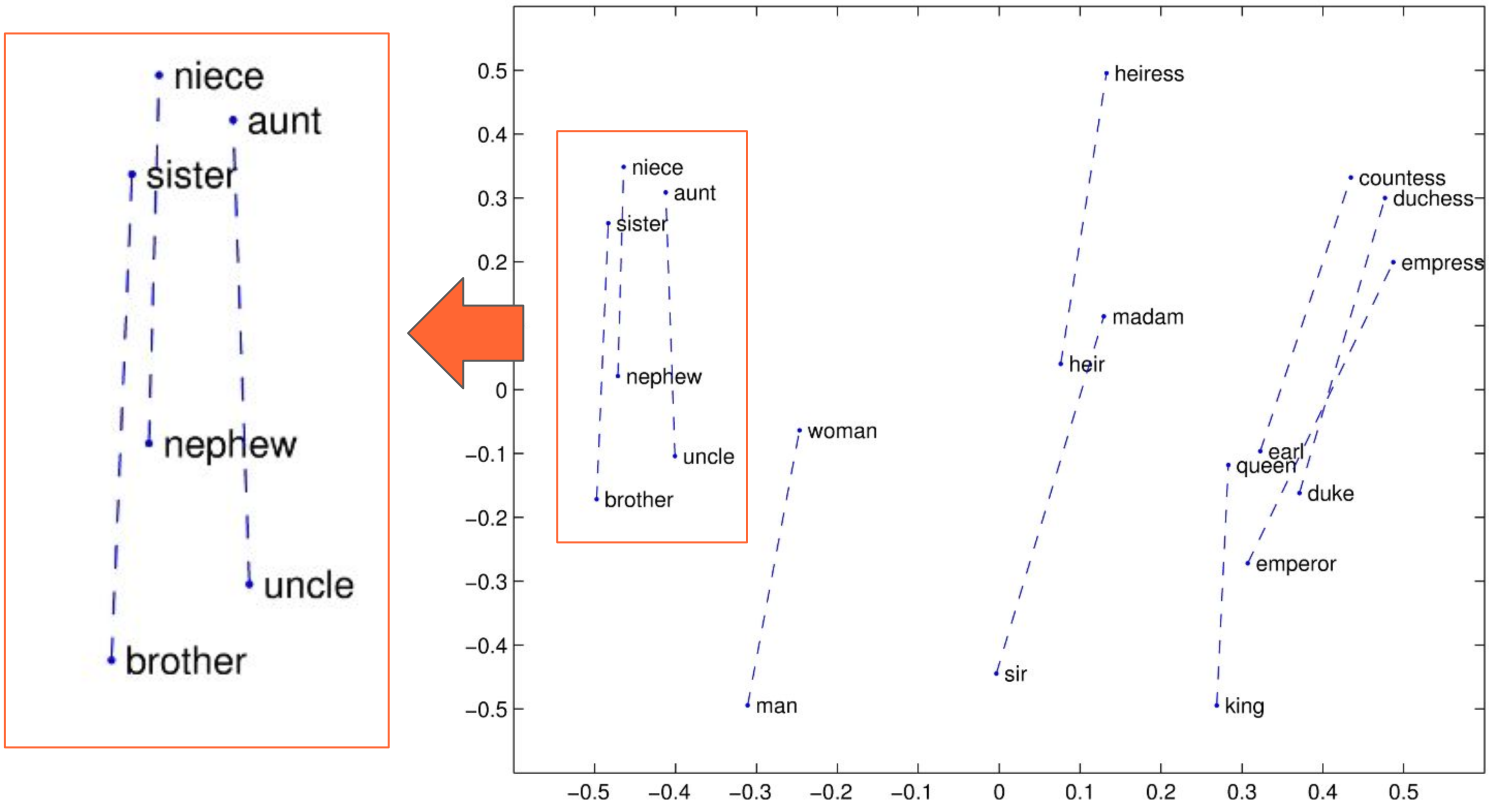


Figure: Christopher Olah, Visualizing Representations

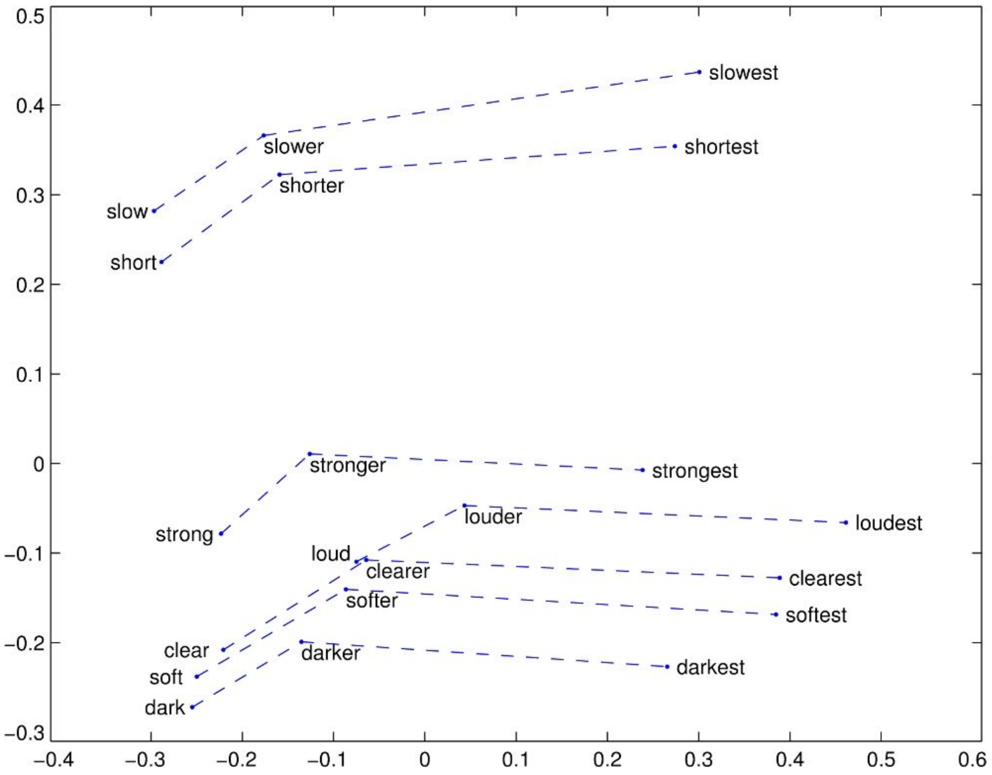
Projection to a word embedding



Man-Woman

Pennington, Jeffrey, Richard Socher, and Christopher Manning. ["Glove: Global vectors for word representation."]() EMNLP 2014

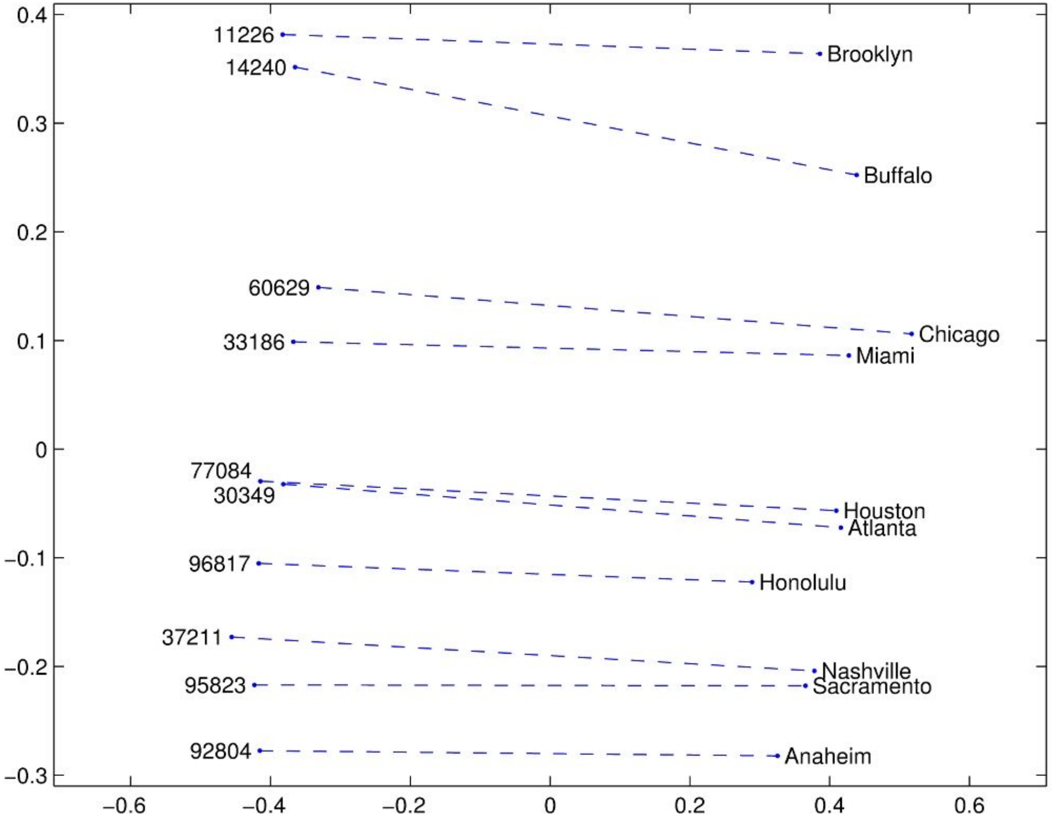
Projection to a word embedding



Comparative-superlative

Pennington, Jeffrey, Richard Socher, and Christopher Manning. ["Glove: Global vectors for word representation."]() EMNLP 2014

Projection to a word embedding



City-Zipcode

Pennington, Jeffrey, Richard Socher, and Christopher Manning. ["Glove: Global vectors for word representation."]() EMNLP 2014

Projection to a word embedding

- Represent words using vectors of dimension *d* (~100 - 500)
- Meaningful (semantic, syntactic) distances
- Dominant research topic in last years in NLP conferences.
- *Good* embeddings are useful for *many* other tasks

Training Word Embeddings

Next word prediction

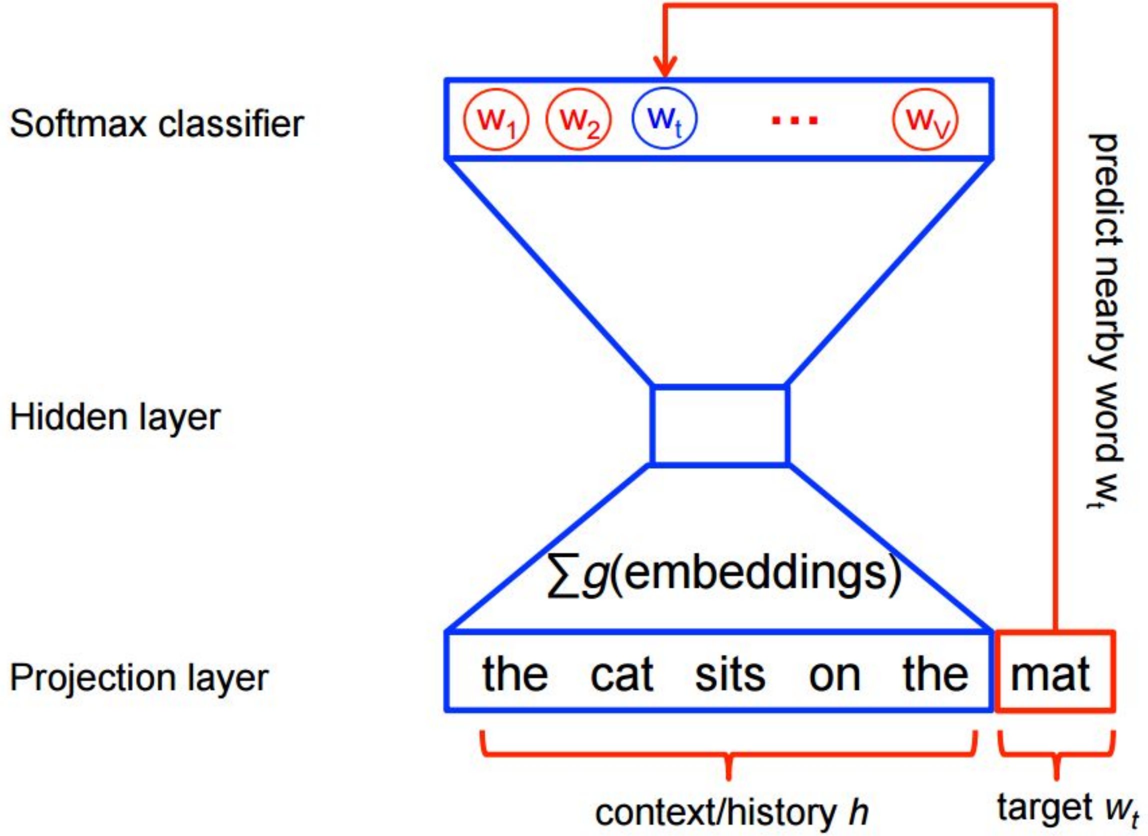
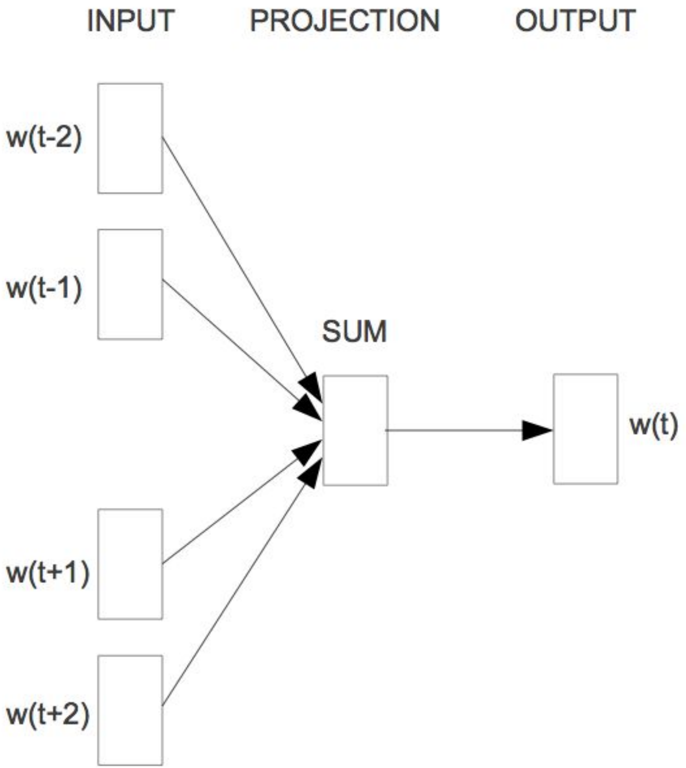


Figure:
TensorFlow tutorial

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. "A neural probabilistic language model." Journal of machine learning research 3, no. Feb (2003): 1137-1155.

Training Word Embeddings

Word2Vec: Continuous Bag of Words



`the cat climbed a tree`

Given context:

`a, cat, the, tree`

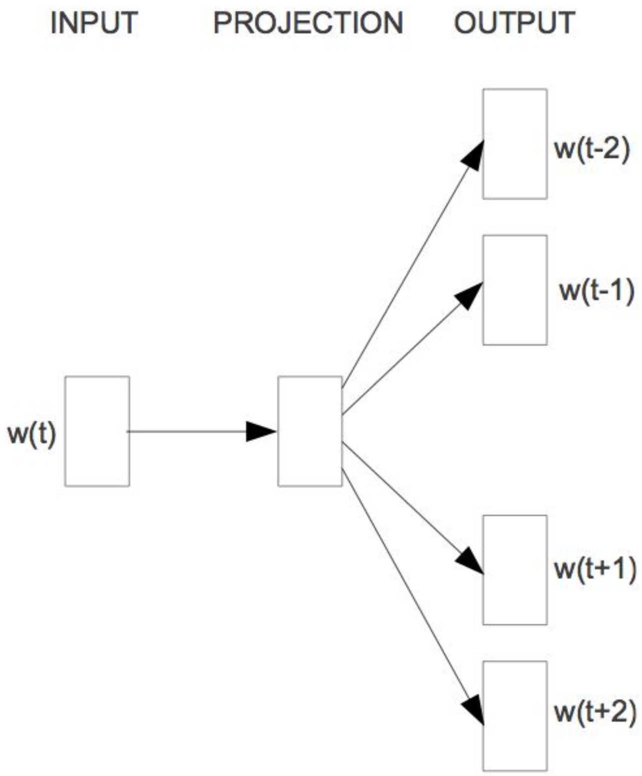
Estimate prob. of

`climbed`

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. ["Distributed representations of words and phrases and their compositionality."]() NIPS 2013

Training Word Embeddings

Word2Vec:
Skip-gram



`the cat climbed a tree`

Given word:

`climbed`

Estimate prob. of context words:

`a, cat, the, tree`

(It selects randomly the context length, till max of 10 left + 10 right)

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

Training Word Embeddings

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

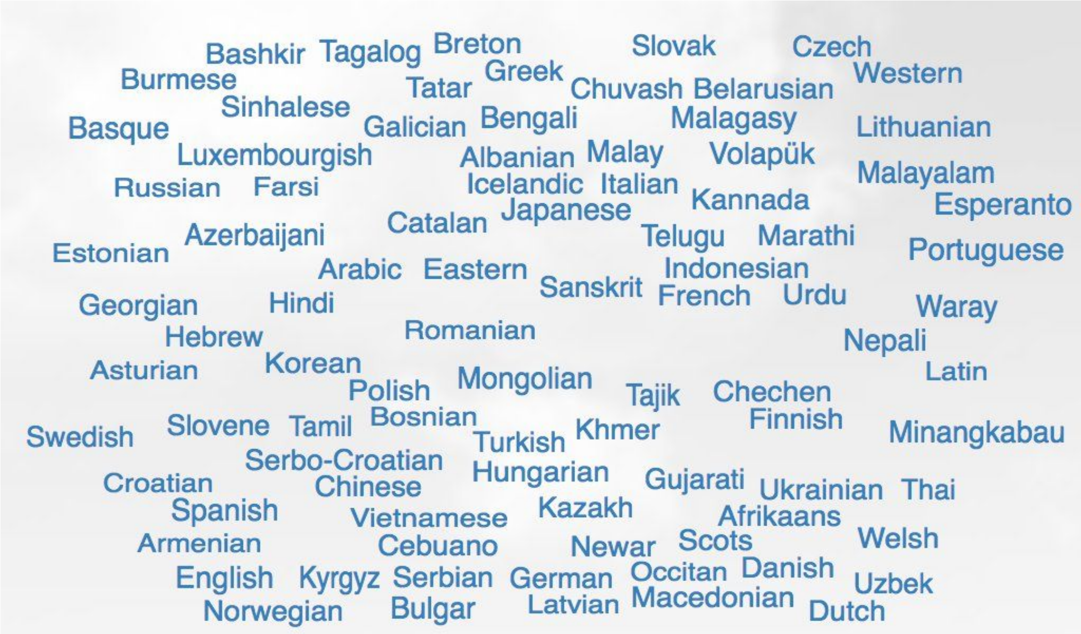
Based on non-zero co-occurrence matrix

The ratio of probabilities encodes some crude form of meaning associated with the abstract concept of thermodynamic phase.

Pennington, Jeffrey, Richard Socher, and Christopher Manning. "Glove: Global vectors for word representation." EMNLP 2014

Multilingual Embeddings

Pre-trained Word Embeddings for 90 languages trained using FastText, on Wikipedia.



facebook research

Questions?