Inspire…Educate…Transform.

Word embeddings

Dr. Kishore Reddy Konda

Mentor, International School of Engineering

Text processing

Words and sentences are of varying length, is this a problem for using them as input to a machine learning algorithm?

Images have pixel intensities which can act as direct inputs to a neural network.

While text needs to be encoded into a vector form.

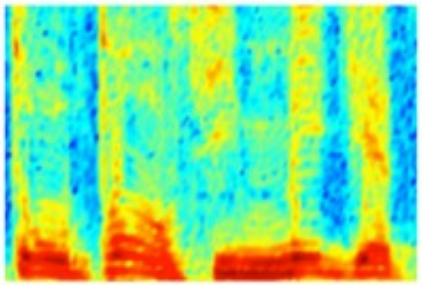
Popular methods for generating word embeddings:

Word2Vec - Mikolov, Tomas; et al. "Efficient Estimation of Word Representations in Vector Space

GloVe - http://nlp.stanford.edu/projects/glove/

Text processing

AUDIO



Audio Spectrogram

DENSE

IMAGES

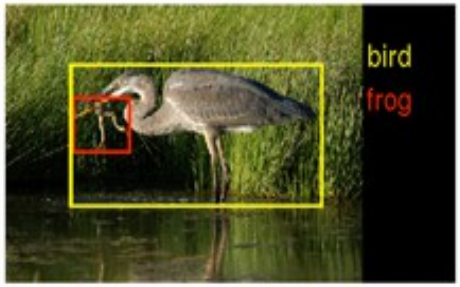


Image pixels

DENSE

TEXT

0	0	0	0.2	0	0.7	0	0	0	...	...

Word, context, or document vectors

SPARSE

Learning word embeddings

Word2Vec and GloVe

- Unsupervised learning neural network based algorithms for obtaining vector representations for words.
- Trained on large corpus of text data.
- representations showcase interesting linear substructures of the word vector space.
- Generates a fixed length vector embedding for each word.

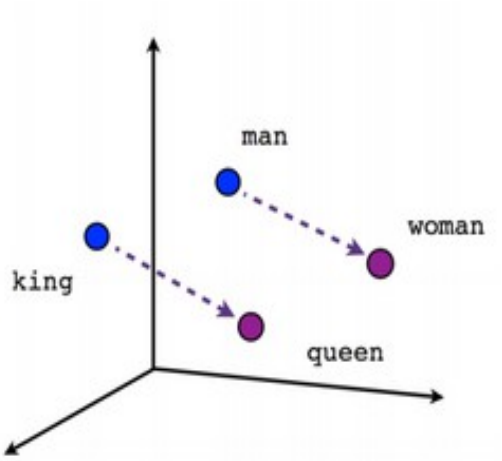
If the length of a given sentence is s, then the dimensionality of the sentence matrix is s×d (where d is the word2vec dimensions).

The parameter d can be in range of 100 to 1000, typically. This is decided when training the unsupervised models (*Word2Vec or GloVe*).

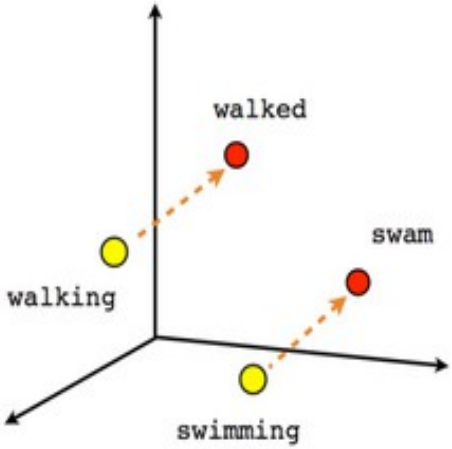
Word2Vec

- Word2vec is a group of related models that are used to produce word embeddings.
- Word2vec was created by a team of researchers led by Tomas Mikolov at Google.
- Algorithm uses a large amount of text to create high-dimensional (50 to 300 dimensional) representations.
- representations of words capturing relationships between words unaided by external annotations.
- Captures many linguistic regularities,
  vec(‘Rome’) = vec(‘Paris’) – vec(‘France’) + vec(‘Italy’).

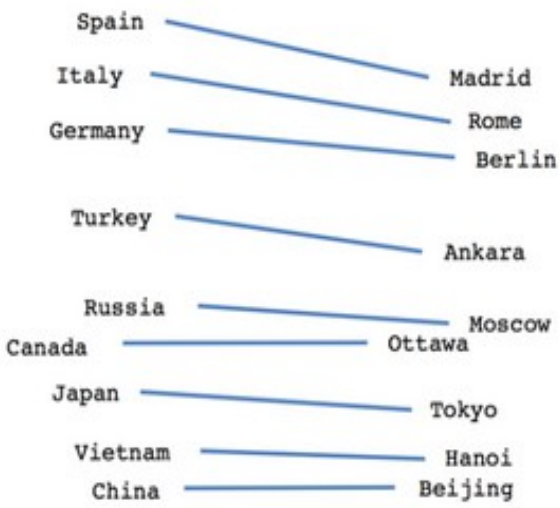
Word2Vec



Male-Female



Verb tense



Country-Capital

Word2Vec: Skip-Gram Model

- A single hidden layer neural network is trained to perform a fake task.
- After training the fake task is dumped and only hidden weights are used.

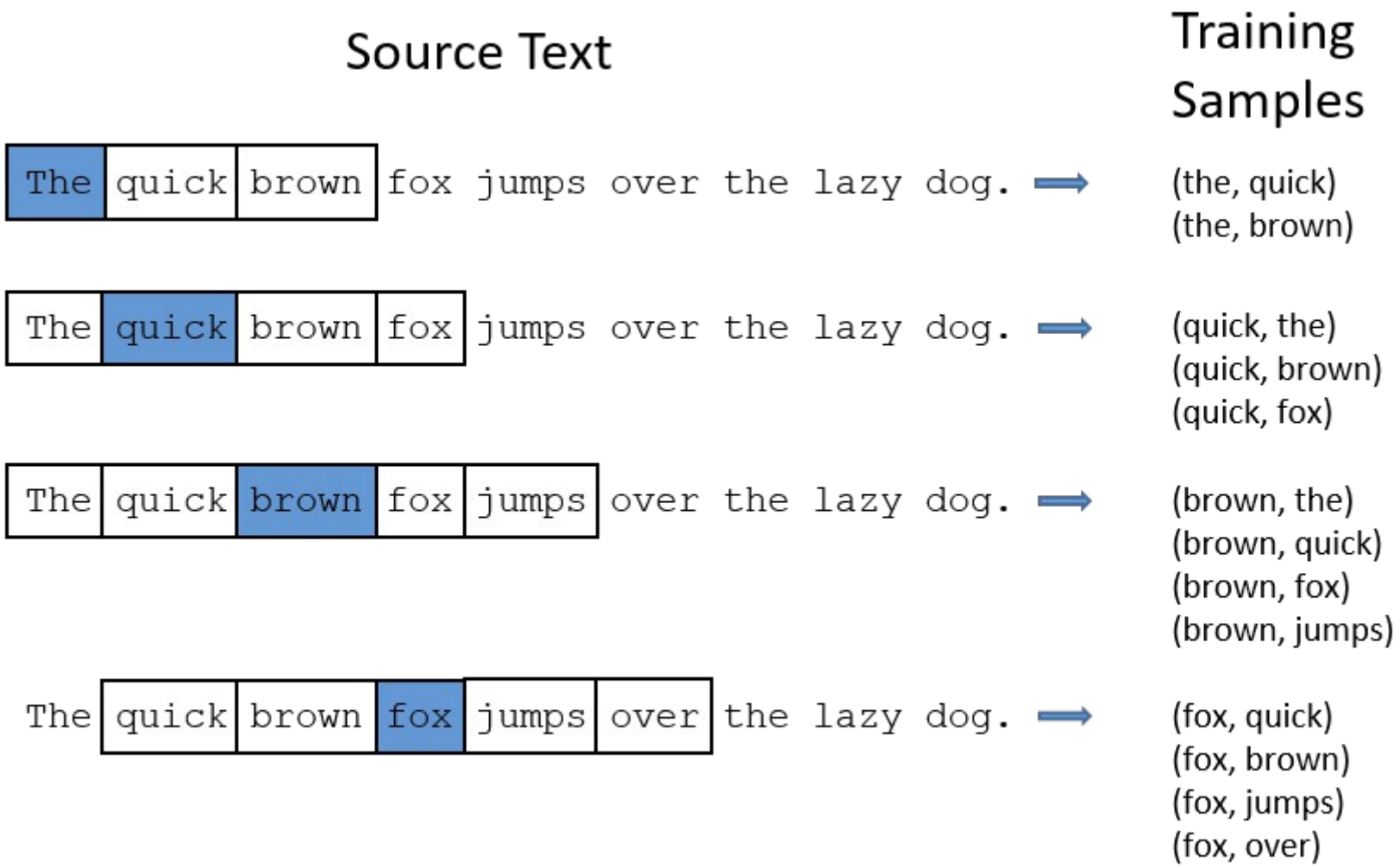
Fake Task: Given a sentence,

"The quick brown fox jumps over the lazy dog."

Predict the probabilities of different words from vocabulary occurring in a fixed window size around the chosen word.

http://mccormickml.com

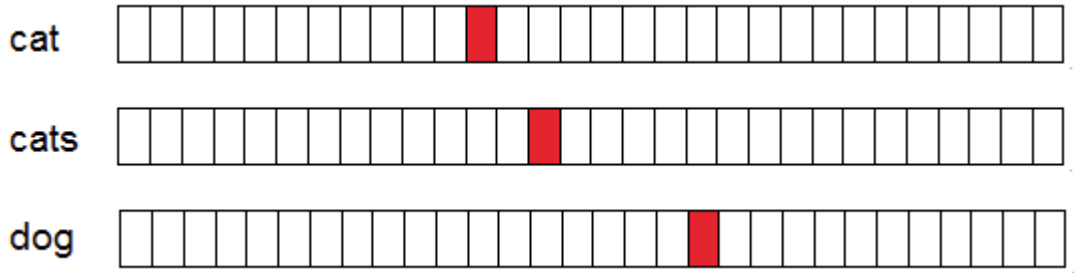
Word2Vec: Skip-Gram Model



http://mccormickml.com

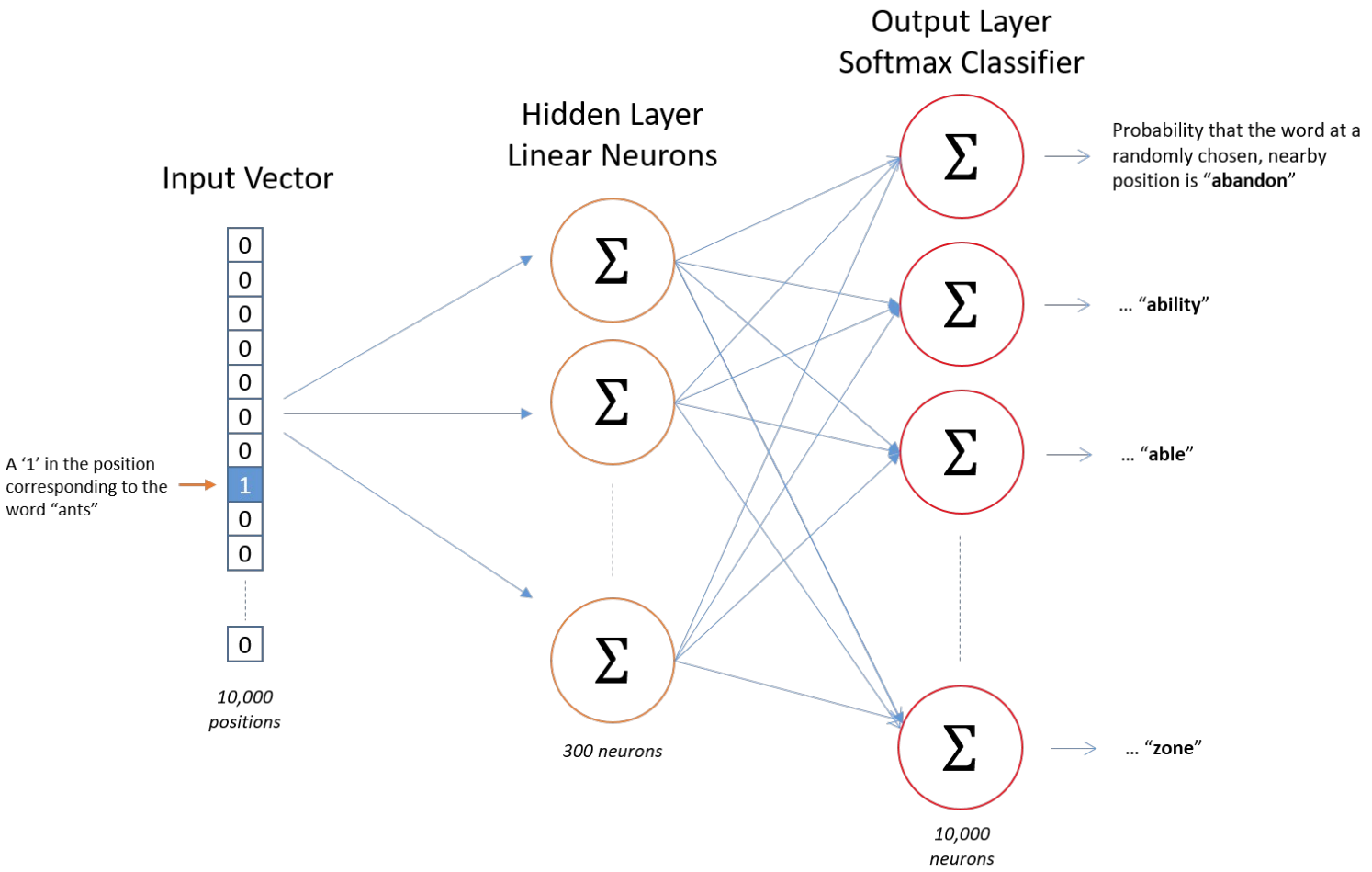
One hot encoding of a word:

- We need a numerical representation for each word to train our skip-gram model.
- If you have a vocabulary of 10000 words treat each word as a state of categorical variable and dummify it.



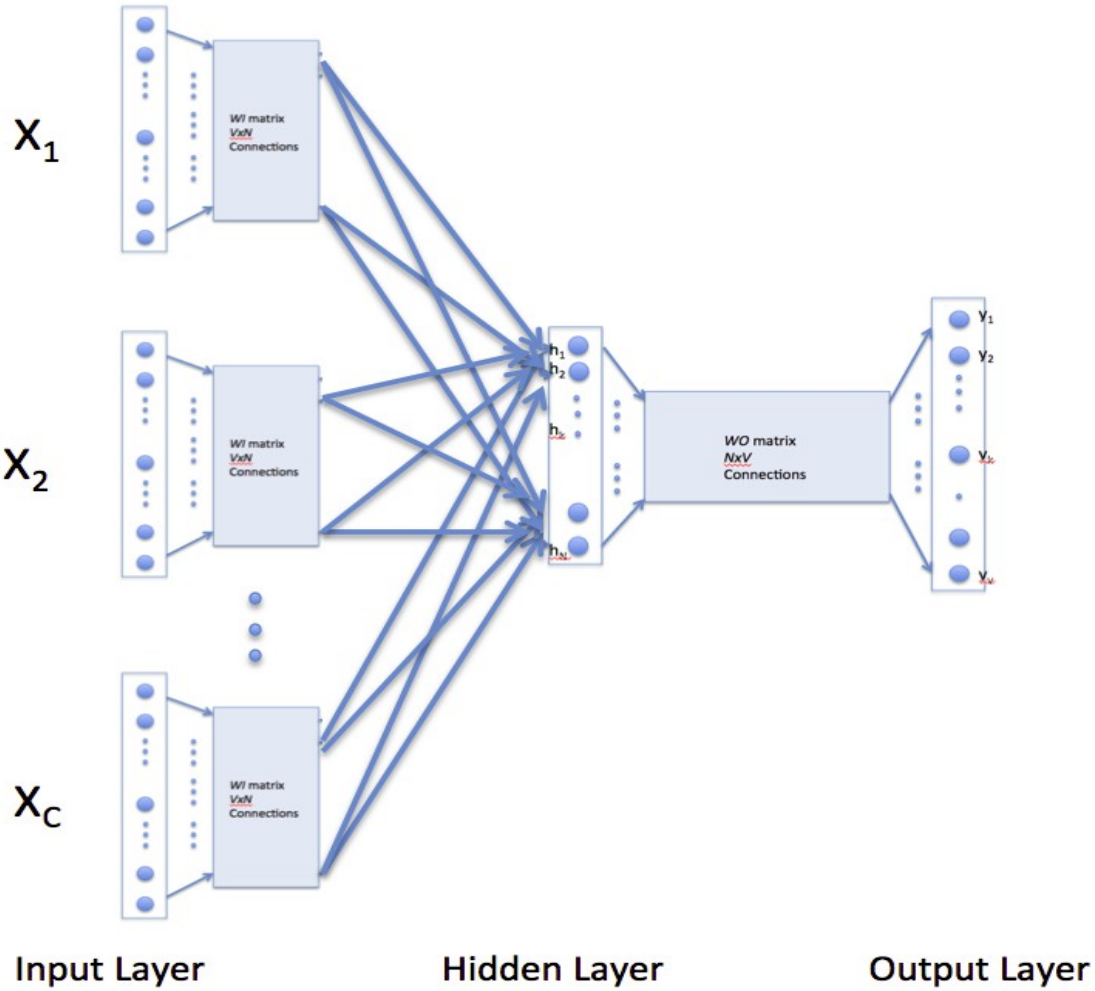
http://mccormickml.com

Word2Vec: Skip-Gram Model



http://mccormickml.com

Word2Vec: CBOW



Training sample:

Given Sentence:

"The quick brown fox jumps over the lazy dog"

(quick,brown,jumps:fox)

(jumps,the,dog: lazy)

http://mccormickml.com

Skip-gram Vs CBOW

- Skip-gram: works well with small amount of the training data, represents well even rare words or phrases.

- CBOW: several times faster to train than the skip-gram, slightly better accuracy for the frequent words

https://code.google.com/archive/p/word2vec/

GloVe: Global Vectors for Word Representation

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014

GloVe

- GloVe is an unsupervised learning algorithm for obtaining vector representations for words.
- Training is performed on aggregated global word-word co-occurrence statistics from a corpus
- Resulting representations showcase interesting linear substructures of the word vector space.

GloVe

- Sometimes, the nearest neighbors according to this metric reveal rare but relevant words that lie outside an average human's vocabulary.

- For example, here are the closest words to the target word frog:

- Frog, frogs, toad, litoria, leptodactylidae, rana, lizard, eleutherodactylus

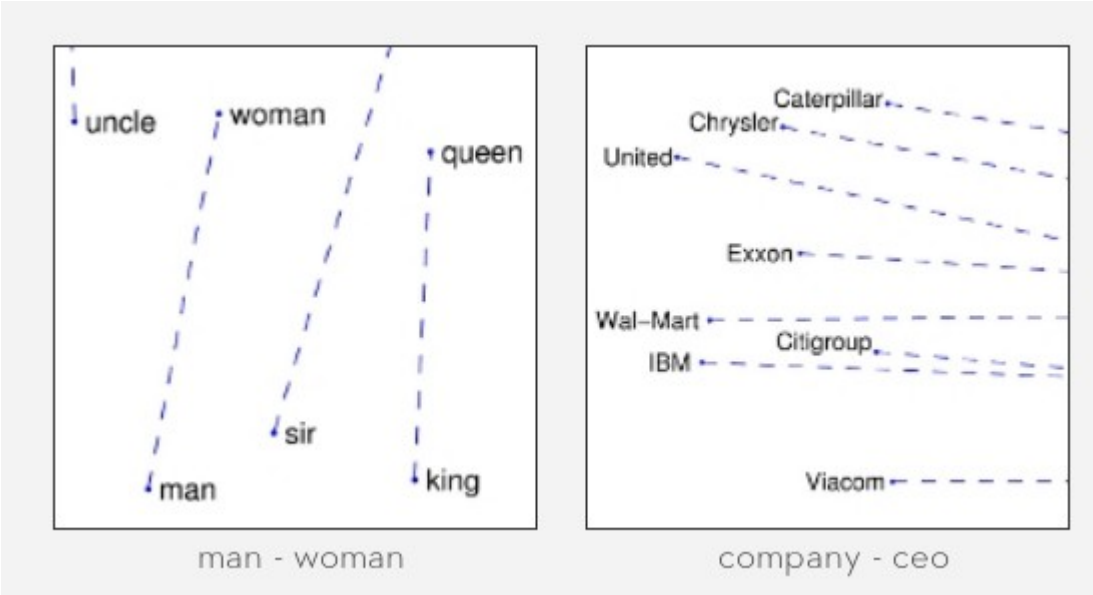
3. litoria

4. leptodactylidae

5. rana

7. eleutherodactylus

GloVe



Linear substructures:The similarity metrics used for nearest neighbor evaluations produce a single scalar that quantifies the relatedness of two words

GloVe: Training

- The GloVe model is trained on the non-zero entries of a global word-word co-occurrence matrix.
- tabulates how frequently words co-occur with one another in a given corpus
- For large corpora, this pass can be computationally expensive, but it is a one-time up-front cost.
- The main intuition underlying the model is the simple observation that ratios of word-word co-occurrence probabilities have the potential for encoding some form of meaning.

GloVe: Training

- The GloVe model is trained on the non-zero entries of a global word-word co-occurrence matrix.
- tabulates how frequently words co-occur with one another in a given corpus
- For large corpora, this pass can be computationally expensive, but it is a one-time up-front cost.
- The main intuition underlying the model is the simple observation that ratios of word-word co-occurrence probabilities have the potential for encoding some form of meaning.

GloVe: Training

- For example, consider the co-occurrence probabilities for target words *ice* and *steam* with various probe words from the vocabulary.

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

GloVe: Training

- For example, consider the co-occurrence probabilities for target words *ice* and *steam* with various probe words from the vocabulary.

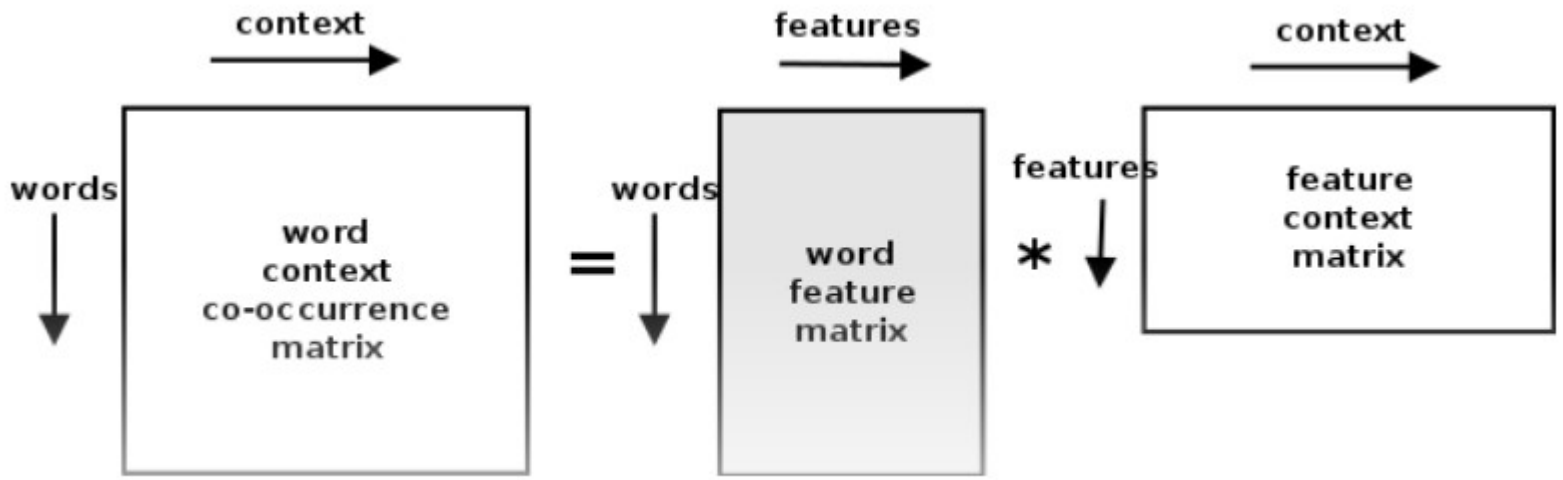
Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

GloVe: Training

- Ice co-occurs more frequently with solid than it does with gas
- steam co-occurs more frequently with gas than it does with solid.
- Both words co-occur with their shared property water frequently.
- Both co-occur with the unrelated word fashion infrequently.

GloVe: Training

- Training objective of GloVe is to learn word vectors such that their dot product equals the logarithm of the words' probability of co-occurrence.



GloVe Vs Word2Vec

- In word2vec, Skip-gram models tries to capture co-occurrence one window at a time

- In Glove it tries to capture the counts of overall statistics how often its appears.

- Both capture linear substructures and tend to perform equally good.