Word Embeddings

Arieda Muço

Central European University

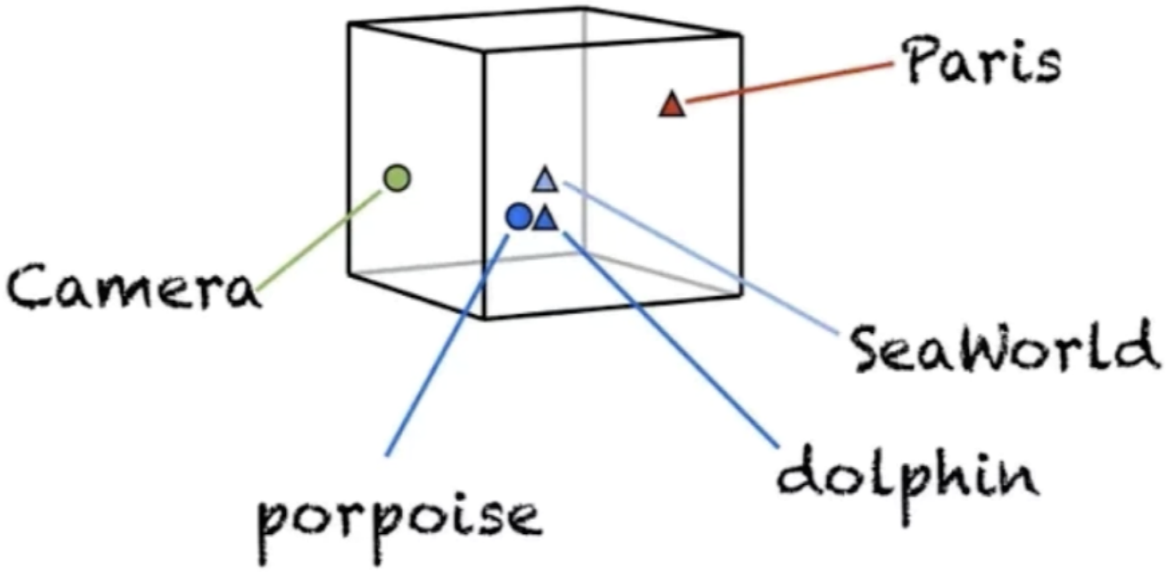
Word Embeddings

- Fancy word, old concept
- Vector representation of a word (we have already seen count-vectorizer, tf-idf)
- What we mean by word embedding is that we are embedding a categorical entity into a vector spacee

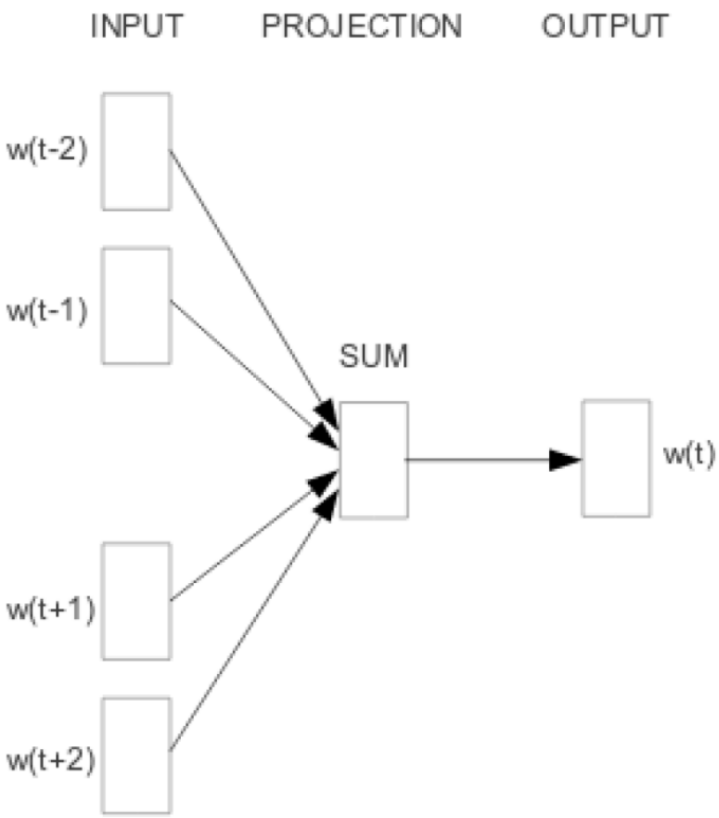
Idea

- Unsupervised extraction of semantics using large corpus (Wikipedia etc)
- Input: one-hot representation of word (as in BoW)
- Use auxiliary task to learn continuous representation

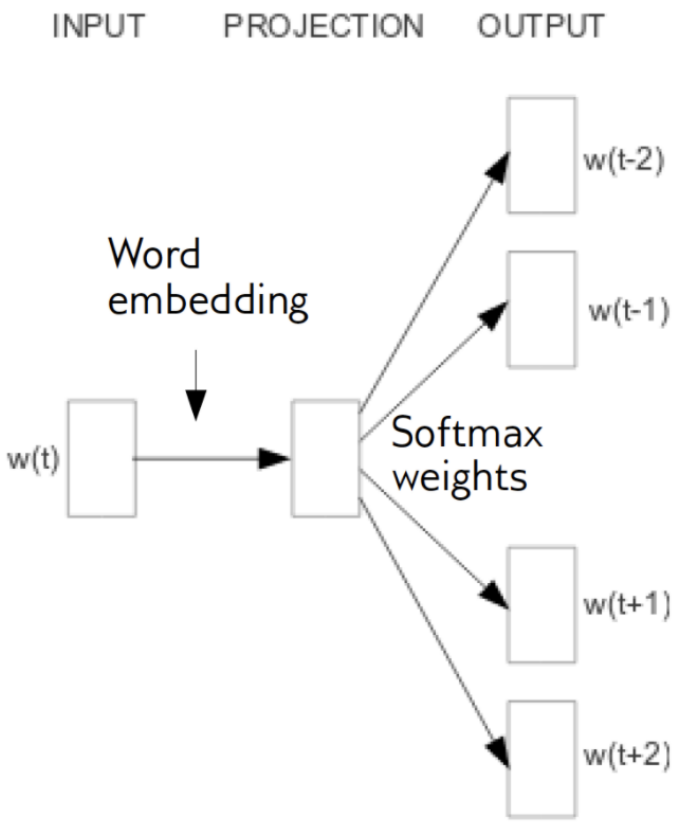
Word Embeddings



Continious Bow vs SkipGram



CBOW



Skip-gram

Word Embeddings

- Word embeddings are dense vector representations of words.
- They capture semantic relationships and contextual meanings.
- Popular algorithms for learning word embeddings include Word2Vec and GloVe.

Durian Example

- Durian is a tropical fruit known for its strong odor.
- In a word embedding space, similar words tend to have similar vector representations.
- Let's say we have a word embedding model that maps words to 4-dimensional vectors.
- The word "durian" might have the following vector representation:

$$\text{durian} = \begin{bmatrix} 0.9 \\ -0.4 \\ 0.5 \\ 0.6 \end{bmatrix}$$

- Other fruits like "mango" and "pineapple" may have similar vector representations.
- This closeness in the embedding space captures the semantic similarity between these fruits.

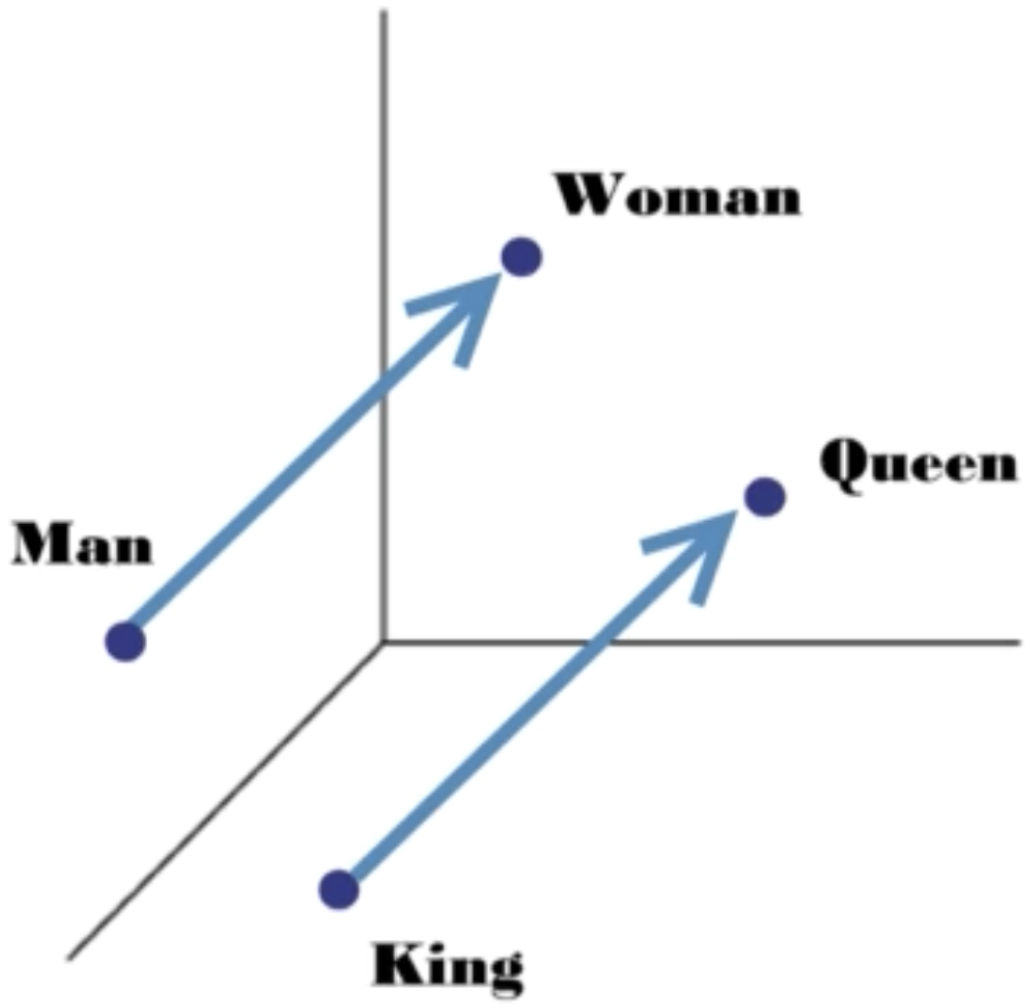
Dense Representation of Words

Dimension	dog	cat	car	house	durian	mango
Dimension 1	-0.2	-0.1	-0.01	-0.01	0.9	0.8
Dimension 2	-0.1	-0.3	0.3	0.4	-0.4	0.7
Dimension 3	0.9	0.9	-0.02	-0.01	0.5	0.5
Dimension 4	-0.2	0.6	0.8	-0.6	0.3	0.3

Table: Sample dense representation of words in a 4-dimensional space

- Each word is represented by a dense vector with multiple dimensions
- Similar words have similar vector patterns across dimensions
- Dense representations enable capturing complex relationships between words

Visualizing Analogies



Examples

- King - Queen $\sim=$ Prince - Princess
- France - Paris $\sim=$ Germany - Berlin
- Japan - Japanese $\sim=$ China - Chinese
- Brother - Sister $\sim=$ Uncle - Aunt
- Walk - Walking $\sim=$ Swim - Swiming