uc3m | Universidad Carlos III de Madrid

HULAT Human Language & Accessibility Technologies

Cutting-edge Deep Learning for NLP learners

Text Representation: word embeddings

Isabel Segura-Bedmar

6 June 2022
Máster Universitario en Inteligencia Artificial, UPM

What is a model language?

- Analyzes the **patterns of human language** for the prediction of words.
  - *Can you please come … ?*
- Probability of a given sequence of words occurring in a text
  - p(“I eat lunch”)
  - p(“I am eating”)
  - p(“Me am eating”)
  - p(“Eating am I”)

Why are important model languages?

- BoW or tf-idf cannot capture semantic information:
  - *Edema de glotis != hinchazón de la laringe*
- Proper language representation is key for developing general-purpose language understanding methods.

How to initialize a neural network?

- Several approaches:
  - random initialization
  - pre-trained word embeddings

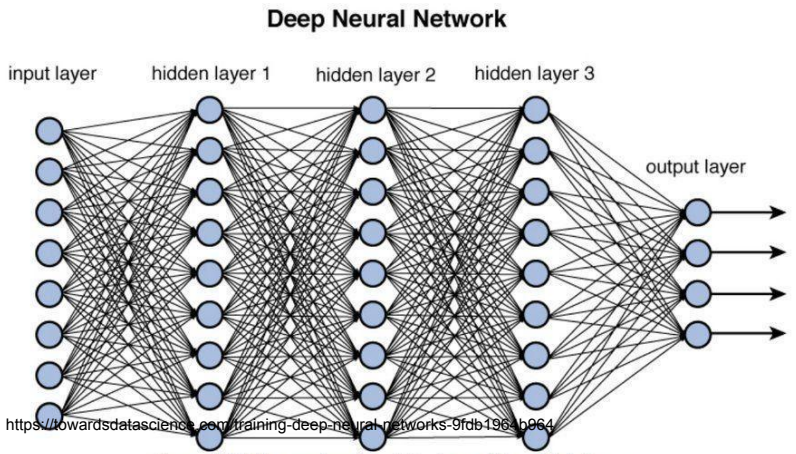


Figure 12.2 Deep network architecture with multiple layers.

Word embeddings (word vectors)

- **Each word** of a vocabulary is represented as a **unique vector (=embedding)**
- **A word embedding model** is a set of **n-dimensional** space vectors.

Neural Networks for Word embeddings

Capture semantic relationships between words.

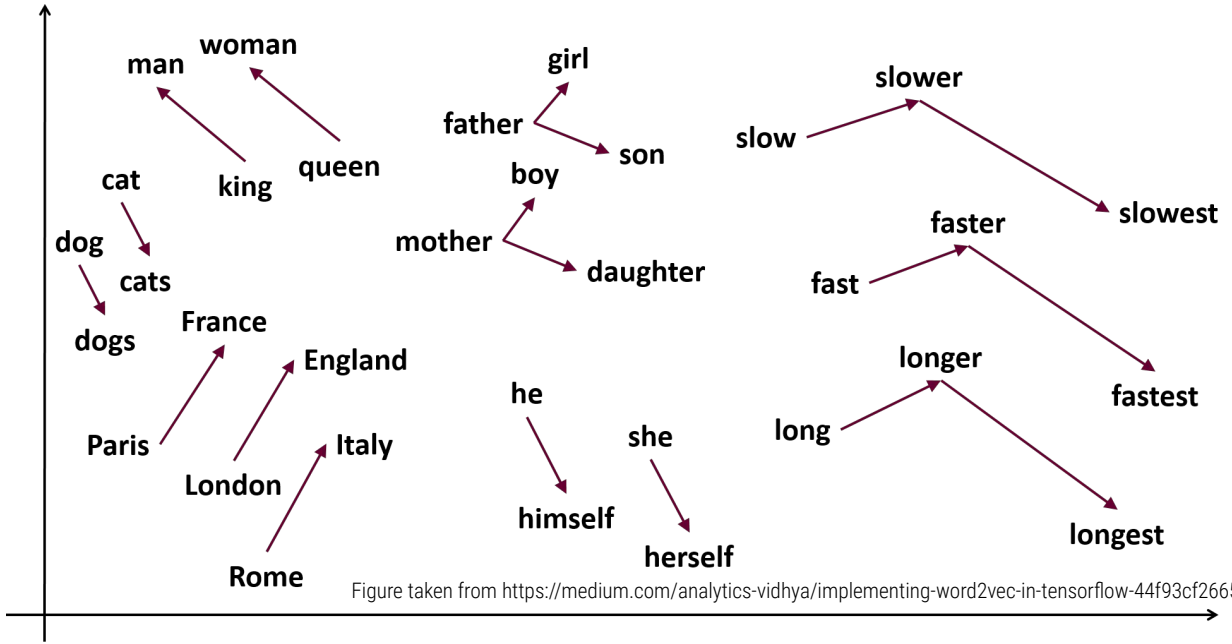


Figure taken from https://medium.com/analytics-vidhya/implementing-word2vec-in-tensorflow-44f93cf2665f

Neural Networks for Word embeddings

Similar words usually happen in similar contexts, so they should have close word embeddings. Cosine distance is very efficient to measure the context similarity.

Distributional Hypothesis
(Harris, 1954)
Words with similar meanings tend to occur in similar context

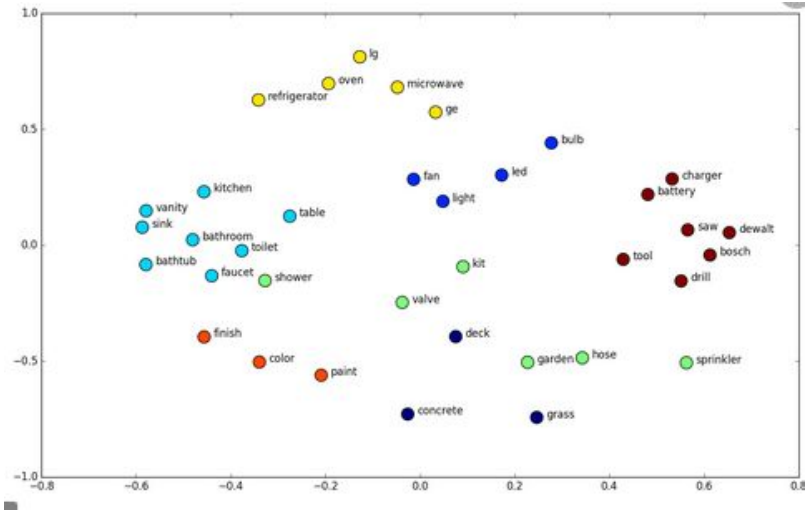


Figure taken from https://laptrinhx.com/word-embeddings-part-2-3902951508/

(Neural) Word embeddings

Allow to **include external knowledge** from the world

(which is not represented in your training data)

(Neural) Word embeddings

For example, if you are developing a Drug NER system, where

- Aspiring exists in the training dataset (labeled as drug), but Ibuprofen does not.

(Neural) Word embeddings

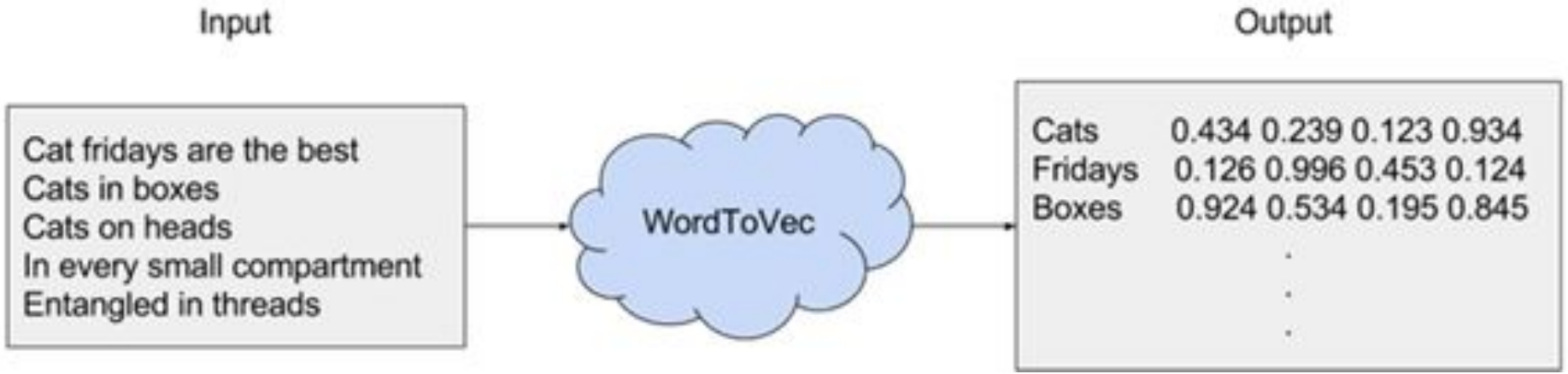
- Aspirin, Ibuprofen should have close vectors in a word embedding model (because both names happen in similar contexts).
  - ***Aspirin is a*** *nonsteroidal anti-inflammatory drug (NSAID)*
  - ***Ibuprofen****, a Nonsteroidal anti-inflammatory Drug, …*
  - ***Ibuprofen*** *is an anti inflammation medicine (a non steroidal anti inflammatory drug or NSAID).*
- Then, the NER system could identify **Ibuprofen** as a drug.

Neural Networks for Word embeddings

- Shallow neural networks trained on a large unlabeled corpus to
  - predict a word based on its context
  - or given a context, to predict the most appropriate word for it).
- **Improve the results** in many **NLP tasks**

Figure taken from https://laptrinhx.com/word-embeddings-part-2-3902951508/

Word2Vec



https://medium.com/@zafaralibagh6/simple-tutorial-on-word-embedding-and-word2vec-43d477624b6d

Word2Vec

- Proposed two architectures to efficiently create word embeddings: continuous **bag-of-words** (CBOW) and **skip-grams** models.
-

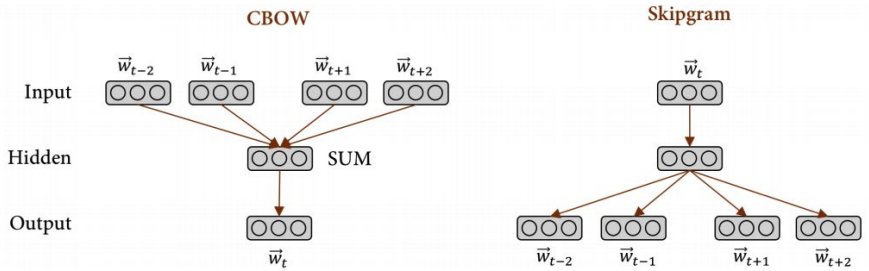


Figure taken from Thomás Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean: Efficient Estimation of Word Representations in Vector Space. ICLR 2013.

Word2Vec: CBOW

CBOW computes the conditional probability of a **target word** given the context words surrounding it across a window of size k.

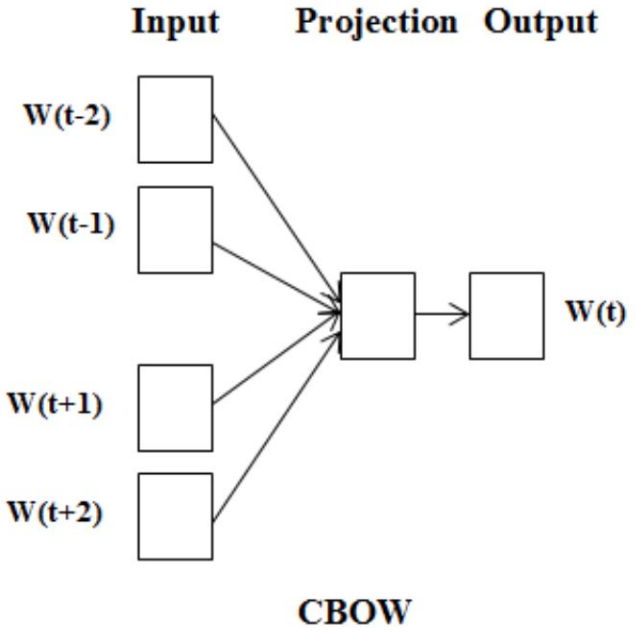
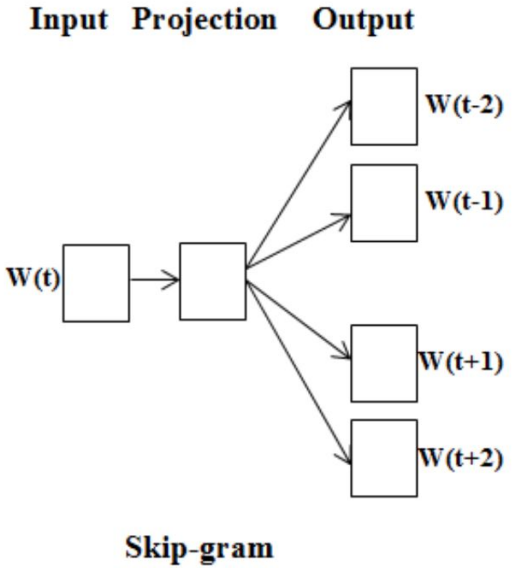


Figure taken from Thomás Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean: Efficient Estimation of Word Representations in Vector Space. ICLR 2013.

Word2Vec: Skip-gram

- **Skip-gram** model **predicts** the surrounding **context words** given the central target word



https://towardsdatascience.com/skip-gram-nlp-context-words-prediction-algorithm-5bbf34f84e0c

Pre-trained word embedding models

- Currently, there are many pre-trained word embedding languages available for research community.
- Spacy already provides its own pre-trained word embedding model.
- Gensim allows us to download different pre-trained word embedding models: https://github.com/RaRe-Technologies/gensim-data
- Repository: http://vectors.nlpl.eu/repository/

Drawbacks of neural word embeddings

- Training word embeddings models require large amount of texts and time.
- Inability to represent phrases (“Joe Biden”, “American Airlines”, “Ford Motor Company”).
- Some words such as *good* and *bad* have very similar word embeddings (very small context window).

Drawbacks of neural word embeddings

Cannot appropriately represent **polysemy words**. However, **ambiguity** is one of the **biggest** challenges in **NLP**

Generate **a single word embedding representation for each word.**

- Work out the *solution* in your head.
- Heat the *solution* to 75° Celsius.

- The *key* broke in the lock.
- The *key* problem was not one of quality but of quantity.

- There are many non-native *pupils* in the class.
- *Pupils*' size changes according to the brightness of light.

http://esl.fis.edu/teachers/support/vocabPoly.htm

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. In NAACL 2019.

Thank you
Question time!!!

isegura@inf.uc3m.es
https://hulat.inf.uc3m.es/nosotros/miembros/isegura
https://github.com/isegura