Applied Artificial Intelligence

Session 21: Language Model, and RNN for modeling sequences (such as text)

Fall 2018

NC State University
Lecturer: Dr. Behnam Kia

Course Website: https://appliedai.wordpress.ncsu.edu/

Text

- Texts are among the most common types of data that we work on in Machine Learning.

Text

- Machine learning algorithms usually do not take the raw text as inputs, rather text must be transformed to numeric values (numeric vectors). We do so by:

  - Segment text into **characters** and transform each to a vector.

  or

  - Segment text into **words** and transform each to a vector.

  or

  - Segment text into **n-grams of words** and transform each to a vector.

Text

- Segment text into **characters** and transform each to a number or a vector.

- Example:

  I am Groot -> “I” “a” “m” “space” “G” “r” “o” “o” “t”

  And then transform each token to a vector

Text

- Segment text into **words** and transform each to a number or a vector.

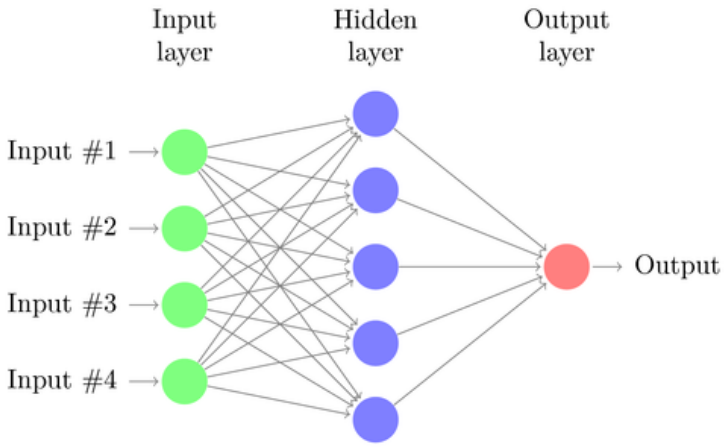
- Example:

  I am Groot -> “I” “am” “Groot”

  And then transform each to a vector

How to model and represent language?

"I" "am" "Groot" ? →



How to transform each token to a vector and represent a sentence with a vector: Method 1: One-hot encoding

Vocabulary={I, am, Groot} ?

One-hot encoding

Vocabulary={I, am, Groot, we, are}

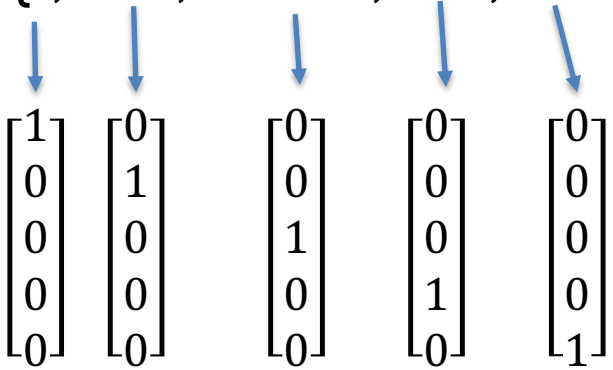
One-hot encoding

Vocabulary={I, am, Groot, we, are}

$$\begin{bmatrix} & \end{bmatrix}$$

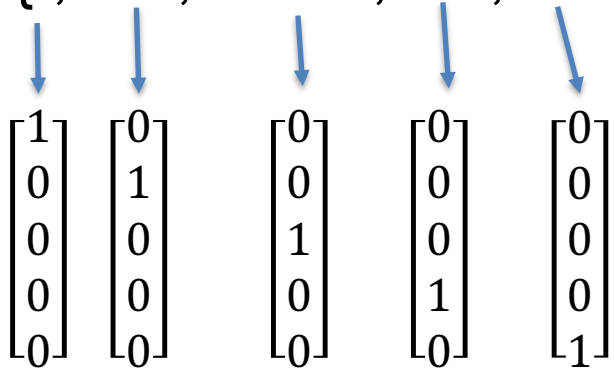
One-hot encoding

Vocabulary={I, am, Groot, we, are}



One-hot encoding

Vocabulary={I, am, Groot, we, are}



I am Groot. → $\begin{bmatrix}1\\1\\1\\0\\0\end{bmatrix}$

We are Groot. → $\begin{bmatrix}0\\0\\1\\1\\1\end{bmatrix}$

Text

- Texts are among the most common types of data that we work on in Machine Learning.

- Segment text into **n-grams of words** and transform each to a a vector.

- Example for 2-grams:

  I am Groot -> “I” “am” “Groot” “I am” “am Groot”

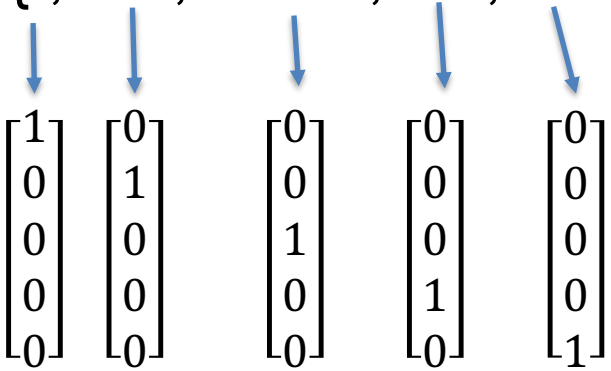
  And then transform each token to a vector

  This representation is called bag-of-2-grams representation of the sentence.

- In most text processing applications words are used as the main tokens.

One-hot encoding: Problems?

Vocabulary={I, am, Groot, we, are}

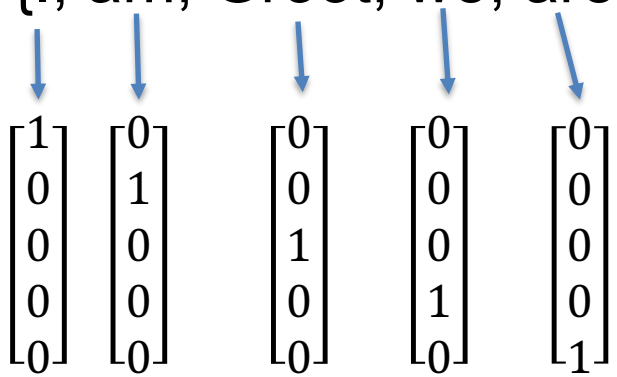


I am Groot. → $\begin{bmatrix}1\\1\\1\\0\\0\end{bmatrix}$

We are Groot. → $\begin{bmatrix}0\\0\\1\\1\\1\end{bmatrix}$

One-hot encoding: Problems?

Vocabulary={I, am, Groot, we, are}



Vocabulary can have 10s of thousands of words.

- Sparse high dimensional vectors.
- Usually hard-coded
- meaningless representation

I am Groot. → $\begin{bmatrix}1\\1\\1\\0\\0\end{bmatrix}$

We are Groot. → $\begin{bmatrix}0\\0\\1\\1\\1\end{bmatrix}$

Word Embedding (word vectorization)

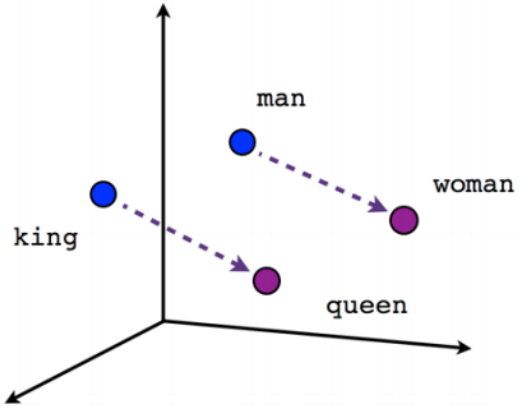
- Each word is represented by a dense, low-dimensional, floating point vectors.
- 256 dimensional, 512 dimensional, or 1024 dimensional.

Word Embedding (word vectorization)

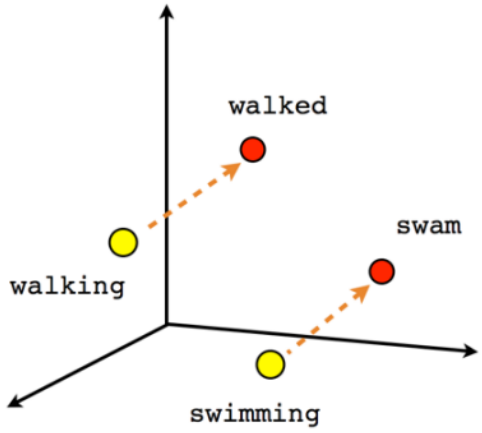
- Each word is represented by a dense, low-dimensional, floating point vectors.
- 256 dimensional, 512 dimensional, or 1024 dimensional.
- Ideally, the geometric relationship between word vectors should reflect the semantic relationship between these vectors.

Word Embedding (word vectorization)

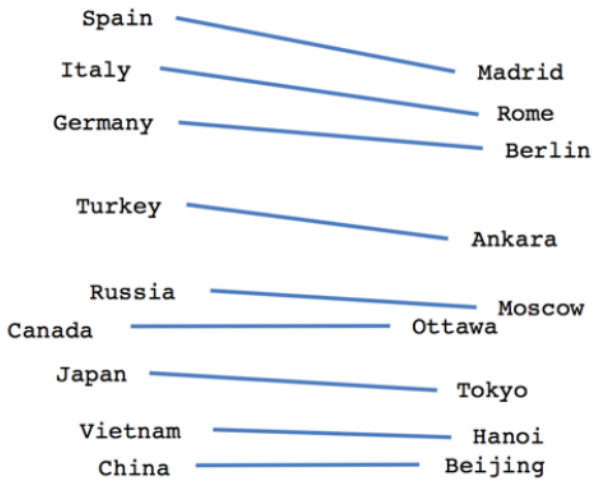
Ideally, the geometric relationship between word vectors should reflect the semantic relationship between these vectors.



Male-Female



Verb tense

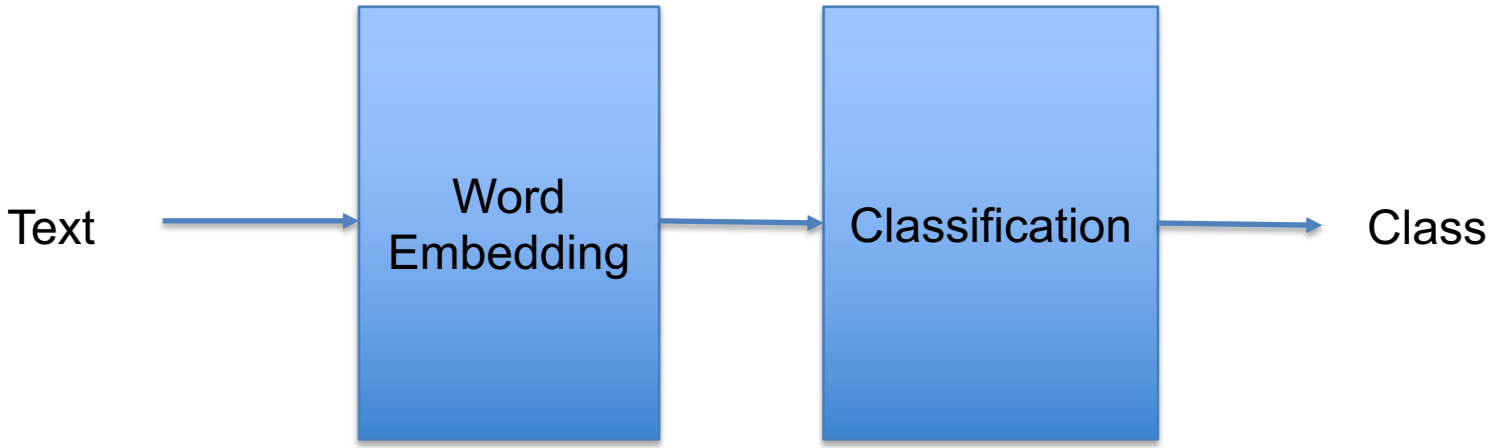


Country-Capital

https://www.tensorflow.org/tutorials/representation/word2vec

How to implement word embedding?

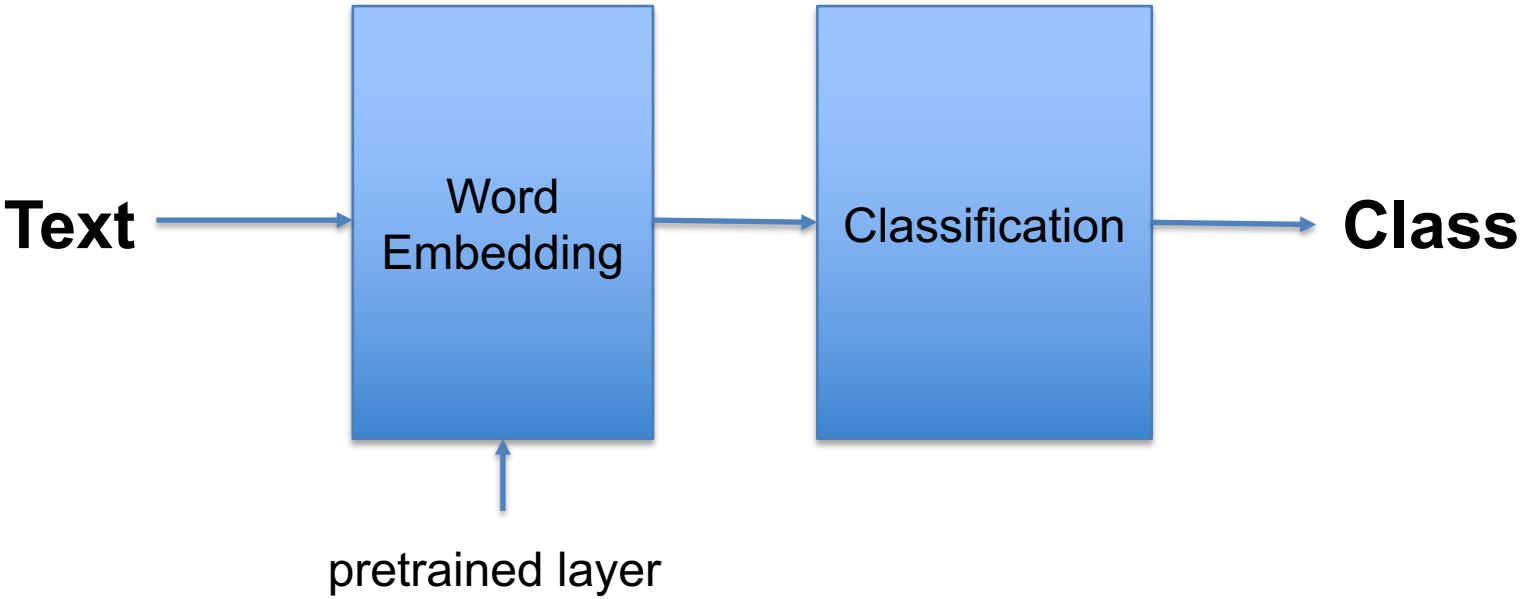
1- Learn jointly with the main task.



How to implement word embedding?

1- Learn jointly with the main task.

2- Load a pretrained word embedding



- Please see the codes: Session 21: Simple RNN for text classification