Simple word2vec

Gerome Yoo

Department of Statistics, Sungkyunkwan University

2018

INDEX

Word Embedding

Example

Korean word2vec
by elnn@elnn.kr 이대근(DAEGEUN LEE)

Try

- 한국 - 서울 + 도쿄 = ?
- 한국 - 제주도 + 대마도 = ?
- 김정은 - 북한 + 한국 = ?

This example shows that by turning the word into vector, we can use basic calculation of "meaning of word" with properties of vector.

WORD EMBEDDING

Word Embedding

Collective name for a set of language modeling and feature learning techniques in natural language processing (NLP) where words or phrases from the vocabulary are mapped to vectors of real numbers.

- *"Word Embedding"* from Wikipedia

$\Rightarrow$ **Simply, way of mapping vectors to words.**

Feature Representations

What is data?

- Data is a set of properties which describes target object.
- Then according to data, model can be built.
- How data express properties effects model performance.
- We call the way how data expresses properties, "Feature Representation"

Feature Representations

For NLP

- Target object is Text and data will be properties of Text.
- What can be the properties of Text?
  First, itself. "Kitty" for the Text "kitty"
  The Length of a word.
  The POS of a word.
  Maybe, where the word is, too.

Feature Representation of Language is to Extract Linguistic Information(such as above) and Represent it.

Feature Representations

Two ways of Linguistic Representation

- Sparse representation ⇒ ex) one-hot encoding
- Dense representation ⇒ ex) word2vec

Sparse representation

- One-hot encoding vector expresses every possible cases with Independent dimension.
- "Sparse" means most elements of vector is Zero, while only few of elements have values.
- Sparse representation is simple, traditional way.

One-hot encoding vector

$$\text{cat} \Rightarrow \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \text{dog} \Rightarrow \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \text{kitty} \Rightarrow \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

One-hot encoding vector

- Most simple way to convert word into a vector
  (1) Score the Words. Then,
  (2) Element of vector which stands for word becomes 1.
  (3) Element elsewhere become Zero.
- If there are N-word, vector will be N-dimensional with only one element with value of 1.

One-hot encoding vector

Drawback of One-hot encoding vector

$$\text{cat} \Rightarrow \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \text{dog} \Rightarrow \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \text{kitty} \Rightarrow \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

- There is no difference between
  relationship of "cat and dog" and "cat and kitty"
- There supposed to be relation between "cat and kitty"
  but there is not, since **"Orthogonal"**.
- It can be concluded that,
  this embedding can not reflect relationship between words.

Feature Representations

Dense Representation

- Dense representation does not represent properties in orthogonal term.
- Instead, represent the word in projection of N-dimension, where N is decided arbitrarily by Conductor.

$$\text{cat} \Rightarrow \begin{bmatrix} 0.7 \\ -0.3 \\ -0.5 \\ 0.42 \\ 0.73 \end{bmatrix}, \quad \text{dog} \Rightarrow \begin{bmatrix} 0.65 \\ 0.4 \\ 0.5 \\ 0.3 \\ 0.71 \end{bmatrix}, \quad \text{kitty} \Rightarrow \begin{bmatrix} 0.7 \\ -0.34 \\ -0.45 \\ 0.4 \\ 0.34 \end{bmatrix}$$

FEATURE REPRESENTATIONS

Dense Representation

- Embedded vector is no more sparse.
  Every element have values.
- Word, "Dense" is used since it is opposite of "Sparse".

Dense Representation = Distributed Representation

- Word, "Distributed" is used to explain the status that
  A word is represented in several dimensions(=attributes).
- In Sparse representation, every element represent independent properties, usually word itself.
- In Dense representation, every element is combined to represent the properties of word.

Feature Representations

Drawbacks of Dense representation

- What exact attributes a dimension stands for can not be known.
- A word is just combination of vector elements.
- The distance between vectors is only index to find out relationship between words.

FEATURE REPRESENTATIONS

Virtue of Dense representation

- Word can be represented with less dimension.

  Sparse representation → Curse of Dimensionality
  Since less dimension, and element value full,
  no sparsity problem.(Ideally)

- Dense representation have more generalization power.

  "Cat and Kitty", in sparse, there are no relation.
  In dense, the distance between vectors will be very close.

Feature Representations

- These advantages work only when word embedding are well trained.
- So how can we learn word embedding?
- There are many types of word embedding.
  Ex) word2vec, GloVe, FastText...etc

After all, Word2Vec is the Topic of this presentation.

IDEA

"You shall know a word by the company it keeps."

- J.R.Firth(1957) -

IDEA

"I like to drink [?]."

(1) Water (2) Wine (3) Food (4) Chair

"I want to learn [?] Language."

(1) Italian (2) French (3) Dynamic (4) Chair

ALGORITHM

2 methods in word2vec

- CBOW : Continuous Bag Of Words
  Use Context to predict word.
- Skip-gram
  Use word to predict context.
- If we flip the CBOW, it becomes the Skip-gram.

CBOW

One-word Context

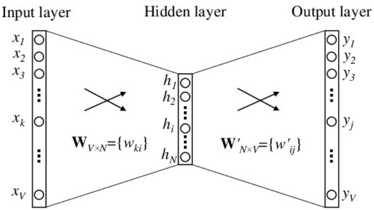

Figure: A Simple CBOW model with only one word in the context

- V : vocabulary size
- N : hidden layer size
- Condition : Adjacent layers are fully connected.
- Input : one-hot encoded vector of given context word
- W' is not transpose of W, is different weight matrix.

CBOW

One-hot encoding vector

- Dimension : $V \times 1$
  Note that using vocabulary $\iff$ no frequnecy info from corpus

CBOW

One-word Context

- V : vocabulary size
- N : hidden layer size
- Adjacent layers are fully connected.
- Input : one-hot encoded vector of given context word

Reference

- word2vec parameter learning explained
- 쉽게 쓰여진 word2vec
- Word2Vec으로 문장분류하기
- 한국어 Word2Vec
- word2vec 관련 이론 정리
- 한글 데이터 머신러닝 및 word2vec을 이용한 유사도 분석