Introduction to learning of text representations

Prakhar Gupta

Department of Computer and Communication Sciences
EPFL

Advanced Topics in Machine Learning

Outline

1. Learning of word representations
   - GloVe
   - Word2Vec

2. Second Main Section
   - Another Subsection

Outline

Matrix Factorization Methods

Matrix Factorization Methods

- Models that can be described as optimization problems of the form

$$\min_{U,V} F(UV^T) \tag{1}$$

GloVe: Global Vectors for Word Representation

Some important definitions and notation

GloVe: Global Vectors for Word Representation

Some important definitions and notation

- Context window is defined as the neighbourhood of a word and its length can be chosen as desired.

GloVe: Global Vectors for Word Representation

Some important definitions and notation

- Context window is defined as the neighbourhood of a word and its length can be chosen as desired.
- Word co-occurence matrix is denoted by X

GloVe: Global Vectors for Word Representation

Some important definitions and notation

- Context window is defined as the neighbourhood of a word and its length can be chosen as desired.
- Word co-occurence matrix is denoted by X
  - where X_{ij} = Number of times word j appears in context of word i

GloVe: Global Vectors for Word Representation

GloVe: Global Vectors for Word Representation

- The aim is to capture $X_i j$ using the source embeddings u_i and target embeddings v_j

GloVe: Global Vectors for Word Representation

- The aim is to capture $X_i j$ using the source embeddings u_i and target embeddings v_j
- The GloVe problem is thus formulated as

$$\min_{U,V} \sum i,j \in W f(X_{ij})(u_i^T v_j + b_i + c_j + log(X_i j)) \tag{2}$$

where $f(X_{ij})$ is the weight assigned to the source-target pair, b_i and c_j are the biases associated with u_i and v_j respectively and W is the vocabulary.

GloVe: Global Vectors for Word Representation

- The aim is to capture $X_i j$ using the source embeddings u_i and target embeddings v_j
- The GloVe problem is thus formulated as

$$\min_{U,V} \sum i, j \in W f(X_{ij})(u_i^T v_j + b_i + c_j + log(X_i j)) \tag{2}$$

  where $f(X_{ij})$ is the weight assigned to the source-target pair, b_i and c_j are the biases associated with u_i and v_j respectively and W is the vocabulary.
- $f(X_{ij})$ is often chosen to be $(\frac{X_{ij}}{Y})^\alpha$ where $Y = \max_{kl} X_{kl}$.

GloVe: Global Vectors for Word Representation

- The aim is to capture $X_i j$ using the source embeddings u_i and target embeddings v_j
- The GloVe problem is thus formulated as

$$\min_{U,V} \sum i, j \in W f(X_{ij})(u_i^T v_j + b_i + c_j + log(X_i j)) \tag{2}$$

  where $f(X_{ij})$ is the weight assigned to the source-target pair, b_i and c_j are the biases associated with u_i and v_j respectively and W is the vocabulary.
- $f(X_{ij})$ is often chosen to be $(\frac{X_{ij}}{Y})^\alpha$ where $Y = \max_{kl} X_{kl}$.
- Empirically $\alpha = \frac{3}{4}$ gives the best performance.

Outline

Word2Vec: CBOW and Skipgram models

Word2Vec: CBOW and Skipgram models

- Uses two difference architectures

Word2Vec: CBOW and Skipgram models

- Uses two difference architectures
  1. Continuous bag-of-words (CBOW)

Word2Vec: CBOW and Skipgram models

- Uses two difference architectures
  1. Continuous bag-of-words (CBOW)
  2. Continuous Skipgram

Word2Vec:CBOW

Word2Vec:CBOW

Intuition

For the CBOW architecture, the task is to

Word2Vec:CBOW

Intuition

For the CBOW architecture, the task is to

- predict the word w given the context $\mathcal{C}(w)$

Word2Vec:CBOW

Intuition

For the CBOW architecture, the task is to

- predict the word w given the context $\mathcal{C}(w)$

Formulation

Given a sequence of training words $w_1, \ldots, w_n$, the maximum likelihood formulation for the CBOW architecture can be written as

$$\sum_{i=1}^{n} \sum_{w_j \in \mathcal{C}(w_i)} \log p(w_j | w_i) \tag{3}$$

Word2Vec:Skipgram

Word2Vec:Skipgram

Intuition

For the Skipgram architecture, the task is to

Word2Vec:Skipgram

Intuition

For the Skipgram architecture, the task is to

- predict the context $\mathcal{C}(w)$ given the word w

Word2Vec:Skipgram

Intuition

For the Skipgram architecture, the task is to

- predict the context $\mathcal{C}(w)$ given the word w

Formulation

Given a sequence of training words $w_1, \ldots, w_n$, the maximum likelihood formulation for the CBOW architecture can be written as

$$\sum_{i=1}^{n} \sum_{w_j \in \mathcal{C}(w_i)} \log p(w_i | w_j) \quad (4)$$

Outline

Blocks

Block Title

You can also highlight sections of your presentation in a block, with it's own title

Theorem

There are separate environments for theorems, examples, definitions and proofs.

Example

Here is an example of an example block.

Summary

- The first main message of your talk in one or two lines.
- The second main message of your talk in one or two lines.
- Perhaps a third message, but not more than that.
- Outlook
  - Something you haven't solved.
  - Something else you haven't solved.

For Further Reading I

A. Author.
Handbook of Everything.
Some Press, 1990.

S. Someone.
On this and that.
Journal of This and That, 2(1):50–100, 2000.