Home Exercise 07

Viet-Hoa Nguyen
Martikelnummer: 2581721

20-00-0947 Deep Learning für Natural Language Processing
Technische Universität Darmstadt

June 8, 2020

1 Mandatory Paper

- Randomly initialized, untrained sentence embeddings produce usable results because their results can be close and sometimes outperform the results of other trained encoders. For this reason, they can be used as a baseline to be compared against when learning other sentence encoders.
- The Bag of random embedding projections creates a sentence representation as follow: first a single random projection is applied in a standard bag of word model. A matrix is randomly initialized consisting of the projection and dimension of input word embedding, then the values for the matrix are sampled uniformly to form the representation of a sentence.

2 Training word2vec Embeddings

2.1 Setup and Parameters

The effects of the parameters:

- **Size**: The number of dimensions of the embedding, e.g. the length of the dense vector to represent each token.
- **Window**: The maximum distance between a target word and words around target word.
- **Negative**: If set to 0, no negative sampling is used, if set to positive, negative sampling will be used, the value of negative specifies how many noise words should be drown.
- **Cbow**: If set to 1, the model type will be set to be Cbow continuous bag of words.

2.2 Training

- The 5 closest words to the word *man*: woman, girl, mortal, hunter, mighty
- The 5 closest words to the word *woman*: man, girl, lover, pregnant, citizen.

3 Sentence Embeddings for Movie Reviews

3.1 Creating Embeddings for Reviews

- The dimension of the average word averaging is 300. The dimension of the concatenated power mean word is 1200.

- When using a power mean with negative parameter p such as the harmonic mean p = -1 then we have the mean equal to $(\frac{1}{n}\sum_{i=1}^{n} x_i^{-1})^{-1} = \frac{n}{x_1^{-1}+...+x_n^{-1}}$, the mean will be undefined if the denominator equals to zero.

3.2 Embedding Comparison

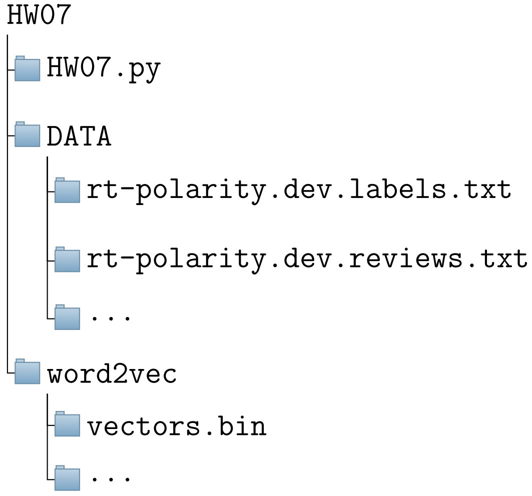
The power means sentence embedding outperforms the naive averaging method in the task hex03 because the different power means can capture more semantics information than the averaging method, e.g. the averaging method ignores word order of a sentence ”your product is easy to use, I do not need any help” has the same sentence embedding as ”I do need help, your product is easy to use”.

In order to run the .py script:

- the word2vec have to train and the binary word vector have to be saved in the file vectors.bin.

- the ”hex07_data.zip” have to be unzip into the DATA folder.

The project folder should have the following structure:



```
HW07
├─HW07.py
├─DATA
│  ├─rt-polarity.dev.labels.txt
│  ├─rt-polarity.dev.reviews.txt
│  └─...
└─word2vec
   ├─vectors.bin
   └─...
```