Minor in AI

SkipGram & GloVe

March 19, 2025

1 Introduction

Word embeddings are dense vector representations of words that capture semantic and syntactic meanings. Two primary approaches to learning word embeddings are:

- Predictive Models (Word2Vec: Skip-gram and CBOW)
- Count-based Models (GloVe)

We will first introduce Skip-gram and then develop an understanding of GloVe.

2 Skip-gram

The Skip-gram model (Mikolov et al., 2013) is a neural network-based approach to learning word embeddings by predicting context words given a target word. The objective is to maximize the probability of context words appearing near a target word in a sentence. GloVe is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space.

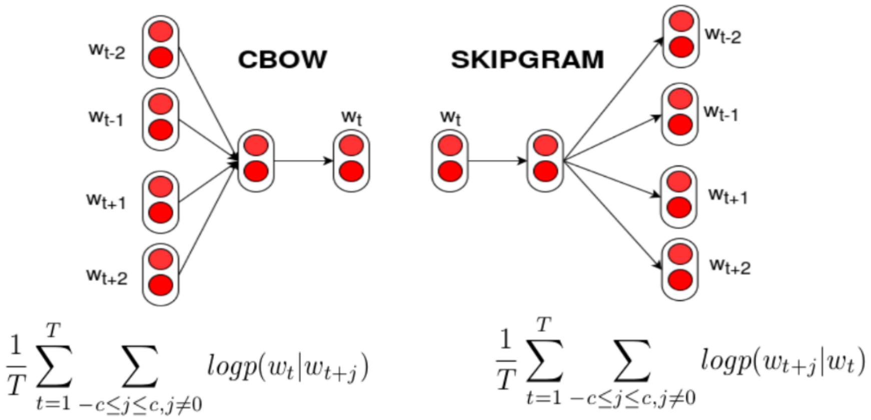
CBOW
SKIPGRAM
w_{t-2}
w_{t-1}
w_{t+1}
w_{t+2}
w_t
$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0} logp(w_t|w_{t+j})$
$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0} logp(w_{t+j}|w_t)$

Figure 1: Word2Vec PC : MLInterview

2.1 How Skip-gram works?

Given a sequence of words $w_1, w_2, \ldots, w_T$, Skip-gram aims to maximize the conditional probability of a context word c given a target word w:

$$\prod_{(w,c)\in D} P(c|w) \tag{1}$$

where D is the set of word-context pairs in the corpus, and $P(c|w)$ is modeled using a softmax function:

$$P(c|w) = \frac{\exp(v_c^T v_w)}{\sum_{c'\in V} \exp(v_{c'}^T v_w)} \tag{2}$$

where:

- v_w is the vector representation of the target word,
- v_c is the vector representation of the context word,
- V is the vocabulary size.

2.2 Example of Skip-gram

Consider the sentence:

”The cat sat on the mat.”

With a context window of size 2, Skip-gram generates the following word-context pairs:

- (cat, the), (cat, sat)
- (sat, cat), (sat, on)
- (on, sat), (on, the)
- (the, on), (the, mat)

The neural network is trained to adjust word vectors so that similar words have similar embeddings.

2.3 Time Complexity of Skip-gram

The denominator in the softmax function involves summing over the vocabulary size $|V|$, leading to a time complexity of $O(|V|)$ per word-context pair. Techniques like negative sampling reduce this to $O(K)$, where K is the number of negative samples.

3 GloVe

GloVe (Global Vectors for Word Representation) is a count-based model that constructs word embeddings by factorizing a word co-occurrence matrix. Unlike Skip-gram, which is predictive, GloVe focuses on capturing the statistical information from the entire corpus. It is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space.

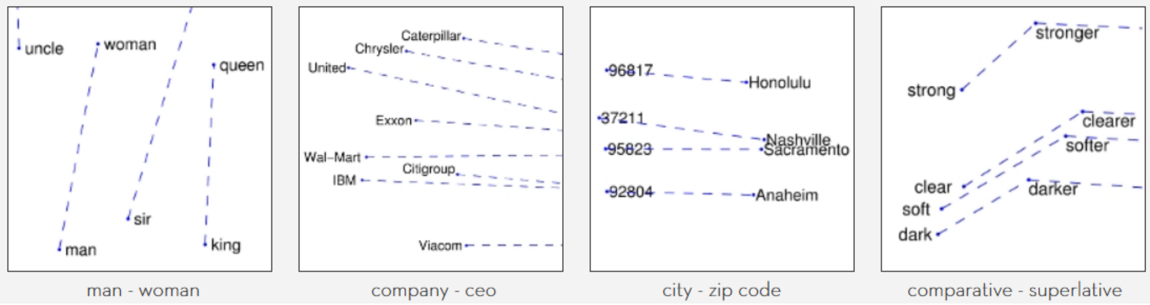

Figure 2: GloVe P.C : Stanford

In order to capture in a quantitative way the nuance necessary to distinguish man from woman, it is necessary for a model to associate more than a single number to the word pair. A natural and simple candidate for an enlarged set of discriminative numbers is the vector difference between the two word vectors. GloVe is designed in order that such vector differences capture as much as possible the meaning specified by the juxtaposition of two words.

The underlying concept that distinguishes man from woman, i.e. sex or gender, may be equivalently specified by various other word pairs, such as king and queen or brother and sister. To state this observation mathematically, we might expect that the vector differences man - woman, king - queen, and brother - sister might all be roughly equal. This property and other interesting patterns can be observed in the above set of visualizations.

3.1 Co-occurrence Matrix

A word co-occurrence matrix X is built, where X_{ij} represents how often word j appears in the context of word i. The key idea is to learn embeddings such that the dot product approximates the log of the ratio of co-occurrences:

$$v_w^T v_c = \log(X_{wc}) \tag{3}$$

3.1.1 Example

Consider the toy corpus:

”The cat sat on the mat.”

We construct a co-occurrence matrix where each entry X_{ij} represents how many times word j appears in the context of word i:

	cat	sat	on	the	mat
cat	0	1	0	1	0
sat	1	0	1	1	0
on	0	1	0	1	1
the	1	1	1	0	1
mat	0	0	1	1	0

Algorithm 1 GloVe Algorithm

```
1. Construct the co-occurrence matrix $X$ from a corpus.
2. Define a cost function that emphasizes word relationships:
       $J = \sum_{i,j} f(X_{ij})(v_i^T v_j + b_i + b_j - \log X_{ij})^2$    (4)
   where $f(X_{ij})$ is a weighting function.
3. Optimize the embeddings $v_w$ using stochastic gradient descent.
```

3.2 Time Complexity of GloVe

The primary computational step is matrix factorization, which is $O(n^2d)$ for a vocabulary of size n and embedding dimension d. However, efficient factorization techniques can reduce this complexity.

4 Key Takeaways

- Skip-gram and GloVe are two primary approaches to learning word embeddings.
- Skip-gram is a predictive model that learns embeddings by maximizing word co-occurrence probabilities.
- GloVe is a count-based model that constructs embeddings through matrix factorization of a word co-occurrence matrix.
- Skip-gram is computationally efficient for large corpora using techniques like negative sampling, while GloVe captures global statistics effectively.
- Choosing between Skip-gram and GloVe depends on the specific requirements of the task and available computational resources.