Deep Learning

Vazgen Mikayelyan

December 29, 2020

FAST DISCOVERING THE FUTURE

Outline

Word2Vec

- We can easily collect very large amounts of unlabeled text data.

Word2Vec

- We can easily collect very large amounts of unlabeled text data.
- Can we learn useful representations (e.g., word embeddings) from unlabeled data?

Bigrams from Unlabeled Data

- Given a corpus, extract a training set $(x_i, y_i)_{i=1}^n$, where $x_i, y_i \in \mathcal{V}$ and $\mathcal{V}$ is the vocabulary.

Bigrams from Unlabeled Data

- Given a corpus, extract a training set $(x_i, y_i)_{i=1}^n$, where $x_i, y_i \in \mathcal{V}$ and $\mathcal{V}$ is the vocabulary.
- For example:

  Hispaniola quickly became an important base from which Spain expanded its empire into the rest of the Western Hemisphere.

  Given a window size of $+/-$ 3, for $x =$ base we get the pairs

  (base, became), (base, an), (base, important),
  (base, from), (base, which), (base, Spain).

The Skip-gram Model

- The training objective of the Skip-gram model is to find word representations that are useful for predicting the surrounding words in a sentence or a document.

The Skip-gram Model

- The training objective of the Skip-gram model is to find word representations that are useful for predicting the surrounding words in a sentence or a document.
- More formally, given a sequence of training words $w_1, \ldots, w_T$, the objective of the Skip-gram model is to maximize the average log probability

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0} \log p\left(w_{t+j}|w_t\right)$$

where c is the size of the training context.

The Skip-gram Model

- The training objective of the Skip-gram model is to find word representations that are useful for predicting the surrounding words in a sentence or a document.
- More formally, given a sequence of training words $w_1, \ldots, w_T$, the objective of the Skip-gram model is to maximize the average log probability

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0} \log p\left(w_{t+j}|w_t\right)$$

  where c is the size of the training context.
- We model $p\left(w_{t+j}|w_t\right)$ using the softmax function:

$$p\left(w_O|w_I\right) = \frac{\exp\left(v'^T_{w_O} v_{w_I}\right)}{\sum_{w=1}^{W} \exp\left(v'^T_{w} v_{w_I}\right)},$$

  where v_w and v'_w are the "input" and "output" vector representations of w, and W is the number of words in the vocabulary.

Negative Sampling

Negative Sampling

- We define Negative sampling by the objective

$$\log \sigma\left(v'^{T}_{w_O} v_{w_I}\right) + \sum_{i=1}^{k} \mathbb{E}_{w_i \sim P_n(w)} \left[\log \sigma\left(-v'^{T}_{w_i} v_{w_I}\right)\right]$$

  which is used to replace every $\log p\left(w_O | w_I\right)$ term in the Skip-gram objective.

Negative Sampling

- We define Negative sampling by the objective

$$\log \sigma\left(v'^{T}_{w_O} v_{w_I}\right) + \sum_{i=1}^{k} \mathbb{E}_{w_i \sim P_n(w)} \left[\log \sigma\left(-v'^{T}_{w_i} v_{w_I}\right)\right]$$

  which is used to replace every $\log p\left(w_O | w_I\right)$ term in the Skip-gram objective.
- Thus the task is to distinguish the target word w_O from draws from the noise distribution $P_n\left(w\right)$ using logistic regression, where there are k negative samples for each data sample.

Negative Sampling

- We define Negative sampling by the objective

$$\log \sigma\left(v'^{T}_{w_O} v_{w_I}\right) + \sum_{i=1}^{k} \mathbb{E}_{w_i \sim P_n(w)} \left[\log \sigma\left(-v'^{T}_{w_i} v_{w_I}\right)\right]$$

  which is used to replace every $\log p\left(w_O | w_I\right)$ term in the Skip-gram objective.
- Thus the task is to distinguish the target word w_O from draws from the noise distribution $P_n\left(w\right)$ using logistic regression, where there are k negative samples for each data sample.
- In the original paper authors chose P_n to be the unigram distribution raised to the 3/4rd power.