Word embeddings

Victor Kitov

v.v.kitov@yandex.ru

Standard word representations

- Denote V=vocabulary size.
- Standard document representations use sparse vectors $x \in \mathbb{R}^V$
  - $x_w = \mathbb{I}[w \text{ occured in the document}]$
  - $x_w = TF_w = \#[w \text{ occured in the document}]$
  - $x_w = TF_w IDF_w$, $IDF_w = \frac{N}{N_w}$
    - N - number of all documents
    - N_w - number of documents, containing w at least once.
- TF and TF*IDF models rely on sparse one-hot word representations [0,0,...0,1,0,...,0]
- V is large, so we want **dense word representations (word embeddings)** $w \rightarrow \mathbb{R}^K$, $K << V$
  - less inputs=>less parameters=>less overfitting, especially for multi-layer perceptron
  - handle synonyms, like "car" and "automobile"

Interpretable word embeddings

- $x \in \mathbb{R}^K$, where x^i is some i-th interpretable feature, e.g.
  - x^1: part of speech
  - x^2: gender (for nouns)
  - x^3: tense (for verbs)
  - x^4: starts from capital letter
  - x^5: #[letters]
  - x^6: category: machine learning, physics, biology, ...
  - x^7: subcategory: supervised, unsupervised, semi-supervised learning
  - ...
- Need to invent features for each task and extract them.
- Want this to be done automatically!

Uninterpretable word embeddings

- Clustering words with similar meaning to similar representations.
- **Distributional hypothesis:**
  words have similar meaning <=> they co-occur together frequently.
- "accuracy of SVM", "SVM gave accuracy", "lower accuracy, compared to SVM"
  - "SVM" and "accuracy" are connected!
- Typical dimensionality of embedding $\in [300, 500]$.

Table of Contents

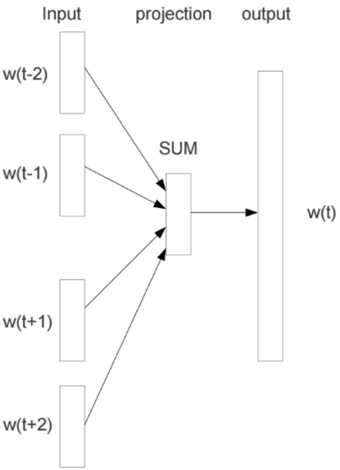
1. Word2vec
   - Models

Word2vec

- Proposed in 2013[1].
- Computationally efficient:
  - Remove computationally expensive hidden layer.
  - Omit expensive denominator calculation.
- Thus can be trained on much bigger datasets.
  - better embeddings, especially for rare words.
- Comments: for each w models evaluate:
  - target word embedding v_w
  - context word embedding $\tilde{v}_w$
- Target&context embeddings may be averaged or concatenated later.

[1] Mikolov et al. (2013), Mikolov et al. (2013)

Continious bag of words (CBOW)



Continuous bag of words (CBOW)

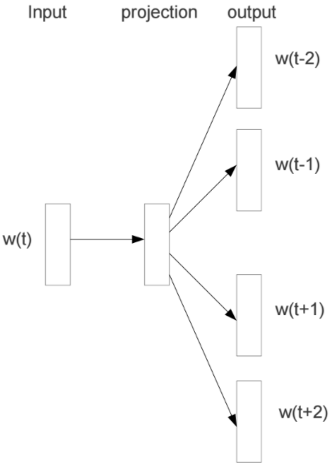
CBOW: predict current word given context.

$$\frac{1}{T}\sum_{t=1}^{T}\ln p(w_t|w_{t-c},..w_{t-1},w_{t+1},...w_{t+c})\to\max_{\theta}$$

where $v_{context}=\sum_{-c\le i\le c,\,i\ne 0}v_{w_{t+i}}$ and

$$p(w_t|w_{t-c},..w_{t-1},w_{t+1},...w_{t+c})=\frac{\exp\left(v_{context}^T\tilde{v}_{w_t}\right)}{\sum_{w=1}^{V}\exp\left(v_{context}^T\tilde{v}_w\right)}$$

Skip-gram model



Skip-gram model

Skip-gram: predict context, given current word:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le i\le c,\, i\ne 0}\ln p(w_{t+i}|w_t)\to\max_{\theta}$$

$$p(w_{t+i}|w_t)=\frac{\exp\left(v_{w_t}^T\tilde{v}_{w_{t+i}}\right)}{\sum_{w=1}^{V}\exp\left(v_{w_t}^T\tilde{v}_{w}\right)}$$