Information Leakage in Embedding Models

Congzheng Song
cs2296@cornell.edu
Cornell University

Ananth Raghunathan
ananthr@cs.stanford.edu
Facebook

CCS 2020

Presentor: Shaofeng Li

October 16, 2020

Overview

Embeddings

A distributed representation on continuous space mapped from the discrete text inputs. Embeddings are pre-trained on a large amount of unlabeled raw data and later fine-tune on downstream tasks with limited labeled data.

Two Granularity:

- Word Embedding: Predict the context word w_i given the center word w_0 by Maximizing the log-likelihood $\log P_V(w_i | w_0)$ where

$$P_V(w_i | w_0) = \frac{exp(\ ^{T}_{w_i} \quad _{w_0})}{\sum_{w \in \{w_i\} \cup V_{neg}} exp(\ ^{T}_{w} \quad _{w_0})}$$

  e.g., Word2Vec, FastText, GloVe

- Sentence Embedding:

  Dual-Encode Architecture

  Lanuage Model: Bert

Sentence Embeddings

Map a variable-length sequence of words x to a x-sized embedding vector $(x) \in \mathbb{R}^d$.

- For a input sequence of l words $x = [w_1, w_2, \ldots, w_l]$ $(x) : X = [\ _1, \ldots, \ _l]$, with a word embedding matrix V.
- Feed X to a LSTM or Transformer and obtain a sequential hidden representation $h = [h_1, h_2, \ldots, h_l]$ for each word in x.
- Reduce the sequential hidden representation to a single vector representation where $(x) = (1/l) \sum_{i=1}^{l} h_i$.

Dual-encoder model

Trained on a pair of context sentences (x_a, x_b), and randomly sampled set of negative sentences X_{neg} that are not in the context of x_a, x_b.

$$P(x_b | x_a, X_{neg}) = \frac{\exp((x_b)^T (x_a))}{\sum_{x \in \{x_b\} \cup X_{neg}} \exp((x)^T (x_a))}$$