Title: Sentence embedding

URL: https://en.wikipedia.org/wiki/Sentence_embedding

PageID: 58348103

Categories: Category:Artificial neural networks, Category:Computational linguistics, Category:Language modeling, Category:Natural language processing

---

Supervised learning

Unsupervised learning

Semi-supervised learning

Self-supervised learning

Reinforcement learning

Meta-learning

Online learning

Batch learning

Curriculum learning

Rule-based learning

Neuro-symbolic AI

Neuromorphic engineering

Quantum machine learning

Classification

Generative modeling

Regression

Clustering

Dimensionality reduction

Density estimation

Anomaly detection

Data cleaning

AutoML

Association rules

Semantic analysis

Structured prediction

Feature engineering

Feature learning

Learning to rank

Grammar induction

Ontology learning

Multimodal learning

Apprenticeship learning

Decision trees

Ensembles Bagging Boosting Random forest

Bagging

Boosting

Random forest

k -NN

Linear regression

Naive Bayes

Artificial neural networks

Logistic regression

Perceptron

Relevance vector machine (RVM)

Support vector machine (SVM)

BIRCH

CURE

Hierarchical

k -means

Fuzzy

Expectation–maximization (EM)

DBSCAN

OPTICS

Mean shift

Factor analysis

CCA

ICA

LDA

NMF

PCA

PGD

t-SNE

SDL

Graphical models Bayes net Conditional random field Hidden Markov

Bayes net

Conditional random field

Hidden Markov

RANSAC

k -NN

Local outlier factor

Isolation forest

Autoencoder

Deep learning

Feedforward neural network

Recurrent neural network LSTM GRU ESN reservoir computing

LSTM

GRU

ESN

reservoir computing

Boltzmann machine Restricted

Restricted

GAN

Diffusion model

SOM

Convolutional neural network U-Net LeNet AlexNet DeepDream

U-Net

LeNet

AlexNet

DeepDream

Neural field Neural radiance field Physics-informed neural networks

Neural radiance field

Physics-informed neural networks

Transformer Vision

Vision

Mamba

Spiking neural network

Memtransistor

Electrochemical RAM (ECRAM)

Q-learning

Policy gradient

SARSA

Temporal difference (TD)

Multi-agent Self-play

Self-play

Active learning

Crowdsourcing

Human-in-the-loop

Mechanistic interpretability

RLHF

Coefficient of determination

Confusion matrix

Learning curve

ROC curve

Kernel machines

Bias–variance tradeoff

Computational learning theory

Empirical risk minimization

Occam learning

PAC learning

Statistical learning

VC theory

Topological deep learning

AAAI

ECML PKDD

NeurIPS

ICML

ICLR

IJCAI

ML

JMLR

Glossary of artificial intelligence

List of datasets for machine-learning research List of datasets in computer vision and image processing

List of datasets in computer vision and image processing

Outline of machine learning

v

t

e

In natural language processing , a sentence embedding is a representation of a sentence as a vector of numbers which encodes meaningful semantic information. [1] [2] [3] [4] [5] [6] [7]

State of the art embeddings are based on the learned hidden layer representation of dedicated sentence transformer models. BERT pioneered an approach involving the use of a dedicated [CLS] token prepended to the beginning of each sentence inputted into the model; the final hidden state vector of this token encodes information about the sentence and can be fine-tuned for use in sentence classification tasks. In practice however, BERT's sentence embedding with the [CLS] token achieves poor performance, often worse than simply averaging non-contextual word embeddings. SBERT later achieved superior sentence embedding performance [8] by fine tuning BERT's [CLS] token embeddings through the usage of a siamese neural network architecture on

the SNLI dataset.

Other approaches are loosely based on the idea of distributional semantics applied to sentences. Skip-Thought trains an encoder-decoder structure for the task of neighboring sentences predictions; this has been shown to achieve worse performance than approaches such as InferSent or SBERT.

An alternative direction is to aggregate word embeddings, such as those returned by Word2vec , into sentence embeddings. The most straightforward approach is to simply compute the average of word vectors, known as continuous bag-of-words (CBOW). [9] However, more elaborate solutions based on word vector quantization have also been proposed. One such approach is the vector of locally aggregated word embeddings (VLAWE), [10] which demonstrated performance improvements in downstream text classification tasks.

Applications

In recent years, sentence embedding has seen a growing level of interest due to its applications in natural language queryable knowledge bases through the usage of vector indexing for semantic search. LangChain for instance utilizes sentence transformers for purposes of indexing documents. In particular, an indexing is generated by generating embeddings for chunks of documents and storing (document chunk, embedding) tuples. Then given a query in natural language, the embedding for the query can be generated. A top k similarity search algorithm is then used between the query embedding and the document chunk embeddings to retrieve the most relevant document chunks as context information for question answering tasks. This approach is also known formally as retrieval-augmented generation [11]

Though not as predominant as BERTScore, sentence embeddings are commonly used for sentence similarity evaluation which sees common use for the task of optimizing a Large language model 's generation parameters is often performed via comparing candidate sentences against reference sentences. By using the cosine-similarity of the sentence embeddings of candidate and reference sentences as the evaluation function, a grid-search algorithm can be utilized to automate hyperparameter optimization [citation needed] .

Evaluation

A way of testing sentence encodings is to apply them on Sentences Involving Compositional Knowledge (SICK) corpus [12] for both entailment (SICK-E) and relatedness (SICK-R).

In [13] the best results are obtained using a BiLSTM network trained on the Stanford Natural Language Inference (SNLI) Corpus . The Pearson correlation coefficient for SICK-R is 0.885 and the result for SICK-E is 86.3. A slight improvement over previous scores is presented in: [14] SICK-R: 0.888 and SICK-E: 87.8 using a concatenation of bidirectional Gated recurrent unit .

See also

Distributional semantics

Word embedding

External links

InferSent sentence embeddings and training code

Universal Sentence Encoder

Learning General Purpose Distributed Sentence Representations via Large Scale Multi-task Learning

References