Improving Sentence Embedding Similarity with Auxiliary Tasks

Jingze Zhang, Gordon Erlebacher
Department of Scientific Computing, Florida State University

Abstract

Learning sentence embeddings is a fundamental problem in Natural Language Processing (NLP), since many of the NLP tasks benefit from an expressive and informative representation. This study improves the quality of sentence embeddings on Semantic Textual Similarity (STS) with two auxiliary tasks: Part of Speech (POS) prediction and Bag of Word (BOW) prediction. We build our model based on SimCSE, a Simple Contrastive Sentence Embedding framework. It predicts the input sentence in a contrastive objective, using dropout layers as a data augmentation strategy. Adversarial training is also applied to disentangle the task specific information into specific channel. Our model slightly outperforms the state-of-the-art sentence embedding models in a preliminary experiment. The model has great potentials to have better performance with careful training. More experiments are in progress.

Methods

▪ Pre-trained models

Language models grow larger and larger with up to hundreds of billions of parameters. Training from scratch is impractical for most researchers. Substantial work has shown that pre-trained models on large corpus can learn general language features which are beneficial for downstream tasks[2]. Finetuning pretrained models on downstream tasks becomes the mainstream approach in NLP. In this work, we initialize our model with the pretrained Roberta[3], providing a warm starting point for faster convergence.

▪ Contrastive Learning

Contrastive learning aims to enhance the embedding quality, such that similar sample pairs stay close to each other in the embedding space, while dissimilar pairs stay apart. In the unsupervised setting, the difficulty lies in how to find similar pairs. SimCSE[1], making use of the dropout layers, obtains the similar pairs by inputting the same sentence twice to the forward pass. We build our model upon the recent successful framework - SimCSE, keep the contrastive objective, and add two auxiliary objectives, as shown in Figure 1.

▪ Adversarial Learning

The adversarial learning concept applied here is slightly different from Generative Adversarial Networks (GANs)[4], which is composed of a generator and a discriminator. The generator learns to generate artificial instances as real as possible, while the discriminator learns to distinguish real or fake instances. In this study, the encoder acts as the generator, and the role of discriminator is to ensure the task specific embeddings only learn current task's knowledge.

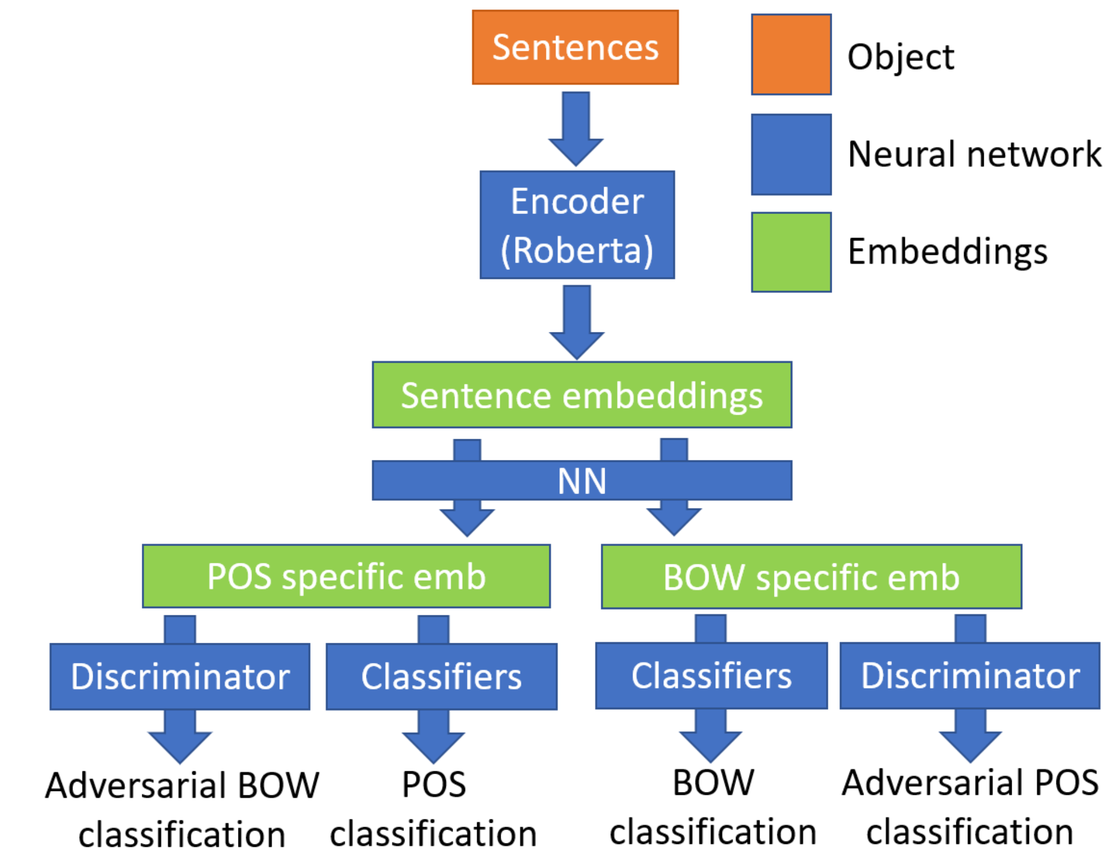

Figure 1. Model architecture. The contrastive learning strategy is applied on the sentence embeddings.

Experiments

▪ Data

1M randomly sampled sentences from English Wikipedia. A pretrained POS tagger is used to get the POS distribution for each sentence.

▪ Evaluation

We evaluate the performance on Semantic Textual Similarity (STS), which measures the degree to which two sentences are semantically equivalent to each other. STS Benchmark[5] comprises a selection of the English datasets used for the STS tasks. Each sentence pair in the task has human annotated scores ranging from 0 for no meaning overlap to 5 for meaning equivalence. Model performance is evaluated by the Spearman correlation between the model produced similarity scores and human judgements.

▪ Training objectives

Classification loss

$$L_{cls}(E, C_{POS}, C_{BOW}) = -\mathbb{E}\left[\log C_{POS}\left(E(x)\right)\right] - \mathbb{E}\left[\log C_{POS}\left(E(x)\right)\right]$$

Adversarial loss

$$L_{adv}(E, D_{POS}, D_{BOW}) = -\mathbb{E}\left[\log D_{POS}\left(E(x)\right)\right] - \mathbb{E}\left[\log D_{BOW}\left(E(x)\right)\right]$$

Contrastive loss (batch level)

$$L_{con}(E) = -\log \frac{e^{sim(E(x_i), E'(x_i))}}{\sum_{j=1}^{B} e^{sim(E(x_i), E'(x_j))}}$$

Notations: E: encoder, C: classifier, D: discriminator, sim: similarity

▪ Results

We conduct our experiments on the STS tasks and compare with the state-of-the-art model SimCSE[1]. Figure 2 compares two models' performance over training iterations on the average of the STS-B and Sick-R tasks. Table 1 shows the evaluation results on seven STS tasks. The highest score for each column is highlighted. Our sentence embedding (Dim=768) improves the STS performance while the task specific embeddings (Dim=640 for BOW task; Dim=128 for POS task) remain relatively high scores.

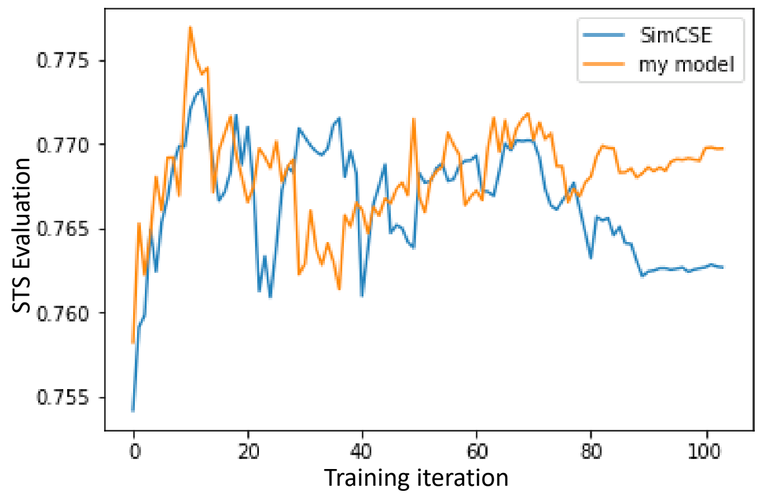

Figure 2. STS evaluation evolution over training iterations.

Table 1. Sentence embedding performance on STS tasks

Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
SimCSE	70.16	81.77	73.24	81.36	80.65	80.22	**68.56**	76.57
$Our_{sentence}$	**70.39**	**82.56**	**73.75**	**83.12**	**81.24**	**81.19**	68.20	**77.21**
Our_{BOW}	69.16	81.45	72.69	81.79	79.98	80.02	67.25	76.05
Our_{POS}	67.70	79.07	71.32	80.34	78.97	78.70	66.44	74.65

Discussion and future works

Our sentence embedding's performance is better but doesn't significantly increase. Task-specific embeddings, especially the POS task embedding, surprisingly have relatively high performance, considering the reduced dimensionalities. More experiments need to be done to search for better hyperparameters.

This model was initially designed for explicitly separating the style and content embeddings. The result provides us a more intuitive understanding in style learning.

Reference

[1] Gao, Tianyu, Xingcheng Yao, and Danqi Chen. "Simcse: Simple contrastive learning of sentence embeddings." arXiv preprint arXiv:2104.08821 (2021).

[2] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[3] Liu, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692 (2019).

[4] Goodfellow, Ian, et al. "Generative adversarial nets." Advances in neural information processing systems 27 (2014).

[5] Cer, Daniel, et al. "Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation." arXiv preprint arXiv:1708.00055 (2017).