SimCSE: Simple Contrastive Learning of Sentence Embeddings

Mohsin Suleman
Riphah International University
Islamabad, Pakistan
mohsinsalman786@yahoo.com

Reproducibility Summary

Scope of Reproducibility
In this reproducibility work, we reproduce Simple Contrastive Learning of Sentence Embeddings. The authors of the original paper claimed that their proposed methodology enhances state-of-the-art sentence embeddings on semantic textual similarity challenges. The main goal of this reproducibility is to again reproduce the results as claimed by the authors of the original paper.

Methodology
We have used the Sentence Transformers library, and implemented Multiple Negatives Ranking Loss for training the SimCSE. In this work we have evaluated the SimCSE on STS benchmark dataset. The framework is evaluated for both supervised and unsupervised learning. We have used distilBERT\textsubscript{base}, BERT\textsubscript{base}, distilRoBERTa\textsubscript{base} and RoBERTa\textsubscript{base} pretrained models of the transformers.

Results
SimCSE can significantly enhance performance on all datasets, with or without additional NLI supervision, beating earlier state-of-the-art models. Our unsupervised SimCSE-BERT\textsubscript{base}, in particular, increases the previous best averaged Spearman's correlation from 72.05 percent to 76.25 percent, making it equivalent to supervised baselines. When NLI datasets are used, SimCSE-BERT\textsubscript{base} improves on the state-of-the-art outcomes to 76.37 percent. The benefits are more obvious on RoBERTa encoders, and our supervised SimCSE with RoBERTa\textsubscript{large} reaches 73.67 percent. Results are shown in Table A.

<table>
<thead>
<tr>
<th>Models</th>
<th>Unsupervised</th>
<th>Supervised</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Spearman</td>
<td>Spearman</td>
</tr>
<tr>
<td>SimCSE-BERT\textsubscript{base}</td>
<td>76.37%</td>
<td>75.67%</td>
</tr>
<tr>
<td>SimCSE- distilBERT\textsubscript{base}</td>
<td>74.09%</td>
<td>72.45%</td>
</tr>
<tr>
<td>SimCSE- distilRoBERTa\textsubscript{base}</td>
<td>70.22%</td>
<td>73.67%</td>
</tr>
<tr>
<td>SimCSE- RoBERTa\textsubscript{base}</td>
<td>69.89%</td>
<td>70.02%</td>
</tr>
</tbody>
</table>

Table A

What was easy
In the reproduction phase, evaluation of the SimCSE tool provided by the authors was easy. We can easily use any model to encode the sentence embeddings. The model also helps to easily compute the cosine similarities between two groups of sentences.
What was difficult

In the reproducibility of the original paper, the running of the authors code was difficult. It consumes a lot of resources. The running of the code takes too much time.

Abstract

SimCSE is a simple contrastive learning system that significantly enhances state-of-the-art sentence embeddings. We begin with an unsupervised technique that takes an input sentence and predicts itself in a contrastive objective using simply ordinary dropout as noise. This basic technique performs remarkably well, matching earlier supervised counterparts. We discover that dropout serves as minimum data augmentation, and that eliminating it causes a representation collapse. Then, using "entailment" pairs as positives and "contradiction" pairings as hard negatives, we present a supervised strategy that includes annotated pairs from natural language inference datasets into our contrastive learning framework. We analyse SimCSE on typical semantic textual similarity (STS) tasks, and our unsupervised and supervised models employing BERTbase achieve an average Spearman’s correlation of 76.3 percent and 81.6 percent, respectively, a 4.2 percent and 2.2 percent improvement over prior best findings. We also show, both theoretically and empirically, that contrastive learning goal regularises the anisotropic space of pre-trained embeddings to make it more uniform, and that it better aligns positive pairings when supervised signals are available.

1. Introduction

Learning universal phrase embeddings is a key challenge in natural language processing that has received substantial research attention [1, 2, 3, 4, 5, 6]. In this paper, we improve cutting-edge sentence embedding approaches and show how a contrastive objective may be particularly effective when combined with pre-trained language models like BERT [7] or RoBERTa [8]. SimCSE is a basic contrastive sentence embedding framework that may generate improved sentence embeddings from unlabelled or labelled data.

Our unsupervised SimCSE predicts the input text using just dropout [9] as noise (Figure 1). In other words, we send the same language to the pre-trained encoder twice: once by using the normal dropout method, and once by using a custom dropout method. We may generate two distinct embeddings by repeating the process again as "positives pairs." Then we look at the other sentences in the model that predicts the positive one among negatives from the same mini-batch as "negatives." Despite its apparent simplicity, this technique excels training objectives like as anticipating what would happen next sentences [5] and discrete by a substantial amount, and even equals earlier data augmentation (e.g., word elimination and replacement) procedures that are overseen. We discovered that dropout functions as a minimum "data augmentation" of hidden representations, but eliminating it results in a representation collapse.

Our supervised SimCSE expands on the previous success of utilising NLI datasets for sentence embeddings [3, 6] by using annotated sentence pairs in contrastive learning (Figure 2). Unlike earlier work, which portrays it as a three-way classification job (entailment, neutral, and contradiction), we take use of the fact that entailment pairs may be utilised as positive cases. We also discovered that adding equivalent contradiction pairs as hard negatives boosts performance much more. When compared to previous approaches that used the same datasets, this straightforward usage of NLI datasets provides a significant improvement. In addition, we
compare NLI datasets to other labelled sentence-pair datasets and discover that NLI datasets are particularly successful at learning phrase embeddings.

To further comprehend SimCSE's high performance, we use Wang and Isola's [10] analytic technique, which measures the quality of learnt embeddings by aligning semantically-related positive pairings and uniformity of the whole representation space. We find that our unsupervised SimCSE increases uniformity while avoiding degraded alignment via dropout noise, hence boosting the expressiveness of the representations. According to the same investigation, the NLI training signal can increase alignment between positive pairings and yield better sentence embeddings. We also make a connection to recent findings that pre-trained word embeddings are anisotropic [11, 12] and demonstrate that, from a spectrum perspective, the contrastive learning objective "flattens" the singular value distribution of the sentence embedding space, thus improving uniformity.

SimCSE is evaluated thoroughly on seven standard semantic textual similarity (STS) tests [13, 14, 15, 16, 17, 18, 19] and seven transfer tasks [4]. On the STS tasks, our unsupervised and supervised models score 76.3 percent and 81.6 percent averaged Spearman's correlation using BERT\textsubscript{base}, respectively, a 4.2 percent and 2.2 percent increase over prior best results. We also outperform the competition in the transfer tasks. Finally, we highlight an incoherent evaluation issue in the literature and synthesise data from several situations for future work in sentence embedding assessment.

2. Related Work

Specifically, work in sentence embeddings extends the distributional hypothesis by predicting adjacent sentences of a given one [1, 2, 5]. [21] demonstrate that merely supplementing the concept of word2vec [22] with n-gram embeddings yields significant results. Several recent (and concurrent) approaches pursue contrastive goals [23, 24, 25, 26, 27, 28, 29] by utilising different perspectives—from data augmentation or different copies of models—of the same sentence or document. In comparison to these works, SimCSE employs the simplest notion, taking various outputs of the same text from a normal dropout, and outperforms them on STS tasks.

Supervised sentence embeddings are expected to outperform their unsupervised counterparts. [3] suggest using NLI datasets to fine-tune a Siamese model, which is then extended to additional encoders or pre-trained models [23, 6]. Furthermore, [31, 32] show that bilingual and back-translation corpora may be used to offer valuable supervision for learning semantic similarity. Another line of research focuses on regularising embeddings to ameliorate the
representation degeneration problem and provides significant improvements over pre-trained language models [12, 33, 34].

3. **Background: Contrastive Learning**

The main ideas of contrastive learning are to pull semantically close neighbours together and push apart non-neighbours [20]. In this work we follow the contrastive framework in Chen [35] and take a cross-entropy objective with in-batch negatives [36, 37].

**Positive instances**

How to build pairings is a fundamental question in contrastive learning. A useful solution in visual representations is to use two random transformations of the same picture (e.g., cropping, flipping, distortion, and rotation) [38]. A similar approach has recently been taken in language representations [39, 40] by utilising augmentation techniques such as word deletion, reordering, and substitution. However, due to the discrete structure of NLP, data augmentation is intrinsically challenging.

In NLP, a similar contrastive learning target has been investigated in many scenarios [37, 41, 42]. Positive pairs are obtained from supervised datasets such as question-passage pairings in these circumstances. Because positive pairs are unique, these techniques always employ a dual-encoder architecture, i.e., two separate encoders for positive pairs. [5] employ contrastive learning with a dual-encoder technique to create current and next sentences for sentence embeddings.

4. **Unsupervised SimCSE**

The concept of unsupervised SimCSE is really simple: we take a set of phrases. The use of separately sampled dropout masks for positive pairings is the key to making this work with identical positive pairs. Dropout masks on fully-connected layers, as well as attention probabilities, are used in typical Transformer training [43]. We just give the encoder the same input twice and receive two embeddings with distinct dropout masks. We see dropout as a simple type of data augmentation: the positive pair uses the same text, and their embeddings differ only in dropout masks. We compare this technique to other STS-B development set training objectives [18].

5. **Supervised SimCSE**

We have shown that introducing dropout noise may maintain a satisfactory alignment for positive pairings. In this part, we investigate if we may use supervised datasets to give stronger training signals for enhancing our approach's alignment. Previous research [4, 6] has shown that supervised natural language inference (NLI) datasets [44, 45] are effective for learning sentence embeddings by predicting whether the relationship between two sentences is entailment, neutral, or contradiction. Instead, we take positive pairings from supervised datasets directly in our contrastive learning framework. We begin by investigating whether supervised datasets are particularly well-suited for creating positive couples. We test a variety of datasets with sentence-pair examples, including 1) QQP4: Quora question pairs; 2) Flickr30k [46]: each image is annotated with 5 human-written captions, and any two captions of the same image are considered a positive pair; 3) ParaNMT [31]: a large-scale back-translation paraphrase dataset5; and 4) NLI datasets [45]. Among all the options, using entailment pairs
from the NLI (SNLI + MNLI) datasets performs the best. We think this is reasonable, as the NLI datasets consist of high-quality and crowd-sourced pairs. Also, human annotators are expected to write the hypotheses manually based on the premises and two sentences tend to have less lexical overlap.

6. Experiment

6.1 Evaluation Setup

We do our studies on seven different semantic textual similarity (STS) challenges. It should be noted that all of our STS experiments are completely unsupervised, and no STS training sets are employed. Even in the case of supervised SimCSE, we merely mean that we use additional labelled datasets for training, as in earlier work [3]. We also assess seven transfer learning tasks. We agree with Reimers & Gurevych [6] that the primary purpose of sentence embeddings is to cluster semantically related phrases and so use STS as the primary outcome.

6.2 Training Details

We begin with pre-trained BERT [7] or RoBERTa [8] checkpoints and use the [CLS] representation as the sentence embeddings. We train unsupervised SimCSE on 106 randomly selected phrases from the English Wikipedia, and supervised SimCSE on a mix of the MNLI and SNLI datasets (314k). The description of the dataset is shown in Table 1. In this work we sample 1 million words from English Wikipedia for unsupervised SimCSE, and the SNLI and MNLI datasets for supervised SimCSE. We have used the Sentence Transformers library, and implemented Multiple Negatives Ranking Loss for training the SimCSE. In this work we have evaluated the SimCSE on STS benchmark dataset. The framework is evaluated for both supervised and unsupervised learning. We have used distilBERTbase, BERTbase, distilRoBERTabase and RoBERTabase pretrained models of the transformers.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>No. of pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsupervised</td>
<td>Phrases from the English Wikipedia</td>
</tr>
<tr>
<td>Supervised</td>
<td>MNLI and SNLI</td>
</tr>
</tbody>
</table>

Table 1: Dataset used for SimCSE

We use the transformers [47] package to implement SimCSE. For supervised SimCSE, we train our models over three epochs, evaluate the model every 250 training steps on the STS-B development set, and save the best checkpoint for final evaluation on test sets. We use the identical procedure for the unsupervised SimCSE, but we only train the model for one epoch. On the STS-B development set, we do a grid-search of batch size of 32, 128 and learning rate of 5e-1 and use the hyperparameter settings. SimCSE is not sensitive to batch sizes as long as the learning rates are tuned appropriately, which contradicts the conclusion that contrastive learning requires large batch sizes [35]. It is most likely owing to the fact that all SimCSE models begin with pre-trained checkpoints, which already give us with a decent starting point.
<table>
<thead>
<tr>
<th></th>
<th>Supervised</th>
<th></th>
<th>Unsupervised</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BERT</td>
<td>RoBERTa</td>
<td>Base</td>
<td>large</td>
</tr>
<tr>
<td>Batch Size</td>
<td>32</td>
<td>32</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>Learning Rate</td>
<td>5e-1</td>
<td>5e-1</td>
<td>5e-1</td>
<td>5e-1</td>
</tr>
</tbody>
</table>

Table 2: Batch sizes and learning rates for SimCSE

6.3 Results

On STS tasks, we compare unsupervised and supervised SimCSE to earlier state-of-the-art sentence embedding algorithms. Average GloVe embeddings [31], average BERT or RoBERTa embeddings, and post-processing approaches such as BERT-flow [12] and BERT whitening are examples of unsupervised baselines [33]. We also compare to several recent methods that use a contrastive objective, such as 1) IS-BERT [23], which maximises the agreement between global and local features; 2) DeCLUTR [24], which uses different spans from the same document as positive pairs; and 3) CT [27], which aligns embeddings of the same sentence from two different encoders. Other supervised techniques using post-processing methods include InferSent [3], Universal Sentence Encoder [31], and SBERT/SRoBERTa [6].

SimCSE can significantly enhance performance on all datasets, with or without additional NLI supervision, beating earlier state-of-the-art models. Our unsupervised SimCSE-BERT\textsubscript{base}, in particular, increases the previous best averaged Spearman's correlation from 72.05 percent to 76.25 percent, making it equivalent to supervised baselines. When NLI datasets are used, SimCSE-BERT\textsubscript{base} improves on the state-of-the-art outcomes to 76.37 percent. The benefits are more obvious on RoBERTa encoders, and our supervised SimCSE with RoBERTa\textsubscript{large} reaches 73.67 percent. Results are shown in Table 3.

<table>
<thead>
<tr>
<th>Models</th>
<th>Unsupervised</th>
<th>Supervised</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Spearman</td>
<td>Spearman</td>
</tr>
<tr>
<td>SimCSE-BERT\textsubscript{base}</td>
<td>76.37%</td>
<td>75.67%</td>
</tr>
<tr>
<td>SimCSE- distilBERT\textsubscript{base}</td>
<td>74.09%</td>
<td>72.45%</td>
</tr>
<tr>
<td>SimCSE- distilRoBERTa\textsubscript{base}</td>
<td>70.22%</td>
<td>73.67%</td>
</tr>
<tr>
<td>SimCSE- RoBERTa\textsubscript{base}</td>
<td>69.89%</td>
<td>70.02%</td>
</tr>
</tbody>
</table>

Table 3: Results on Supervised and Unsupervised Dataset of SimCSE

7. Analysis

In this part, we undertake more analysis to better understand SimCSE's internal workings. In general, models with higher alignment and homogeneity perform better, validating Wang and Isola's findings [10]. We also see that (1) while pre-trained embeddings have good alignment, their uniformity is poor (i.e., the embeddings are highly anisotropic); (2) post-processing methods like BERT-flow and BERT-whitening greatly improve uniformity but suffer from alignment degeneration; (3) unsupervised SimCSE effectively improves uniformity while maintaining good alignment; and (4) incorporating supervised data in SimCSE further improves alignment. Using SBERT\textsubscript{base} and SimCSE-BERT\textsubscript{base}, we do a small-scale retrieval experiment. We utilise 150k captions from the Flickr30k dataset as queries and obtain related
phrases (based on cosine similarity). As numerous examples, the phrases retrieved by SimCSE are of greater quality than those obtained by SBERT.

8. Conclusion

In this work we reproduce SimCSE, a simple contrastive learning framework that significantly improves state-of-the-art sentence embeddings on semantic textual similarity challenges, in this paper. We offer an unsupervised technique that predicts the input sentence itself using dropout noise, as well as a supervised one that uses NLI datasets. We further validate the inner workings of our method by examining SimCSE's alignment and uniformity in comparison to other baseline models. We feel that our contrastive goal, particularly the unsupervised one, has a wider use in NLP. It offers a fresh look at data augmentation using text input, and it may be expanded to other continuous representations and integrated into language model pre-training.

References


Yu Meng, Chenyan Xiong, Payal Bajaj, Saurabh Tiwary, Paul Bennett, Jiawei Han, and Xia Song. 2021. COCO-LM: Correcting and contrasting text sequences for language model pretraining. arXiv preprint arXiv:2102.08473.


