Reproducibility report formatting instructions for ML Reproducibility Challenge 2022

Anonymous Author(s)
Affiliation
Address
email

Reproducibility Summary

Scope of Reproducibility

The study in *SimCSE: Simple Contrastive Learning of Sentence Embeddings* (1) targets to improve sentence embeddings by using contrastive learning and shows their success in Semantic Textual Similarity and Transfer Learning tasks. We reproduced the success by retraining the model using the authors' code and tested in the shared tasks.

Methodology

We used authors' code to reproduce the results. In total 4 models are trained, which each took around 5 hours using GPU. Google Colab environement is used for GPU usage. 8 tests are done in CPU, where each took around 4 and a half hour. Apart from some minor bug fixes, no additional implementation are made to the original code.

Results

We trained 2 of the most successfull models, SimCSE-BERT base and SimCSE-RoBERTa base, and in both test sets, we reproduced the results within 2% error margin in the average scores. However, most of the time our results were lower than the pre-trained models and only a few times we have got better results. For some specific tests, the error margin went up to the 5%. We could not train the most successful model, SimCSE-RoBERTa large, for comparison.

What was easy

Both training and test scripts are easy to use. Parameters are specified clearly and one script is enough to get all relevant results.

What was difficult

Training and testing takes time. Training not taking a day might be an ideal situation where most model trainings take days to complete, but testing taking hours make it difficult to test different kinds of parameters to see how they affect the results.

Communication with original authors

Communication with authors will be made to inform them about bug fixes but it is planned to do after project submission, since then github repository will be made public so that authors can see the code changes more clearly within commit differences.

Submitted to ML Reproducibility Challenge 2020. Do not distribute.

1 Introduction

SimCSE: Simple Contrastive Learning of Sentence Embeddings is a study that focuses on improving sentence embeddings using contrastive learning. It uses pre-trained transformed-based models BERT and RoBERTa, gets sentence representations from these models, and trains its model by using contrastive learning on an encoder. This can be done both by supervised and unsupervised learning.

The model is compared with the base models and improved extended versions of BERT and RoBERTa models. It is also compared with models that only produces vectors, such as GloVe and Universal Sentence Encoder. Tests are done on Semantic Text Similarity (STS) and Transfer Task (TT) test sets.

2 Scope of reproducibility

The paper's fundamental claim is that the study has improved sentence embeddings compared to the state-of-the-art models. The comparison is made on STS and TT tests.

- SimCSE-BERT model performs better on STS tasks compared to the base BERT model and its variations.
- SimCSE-RoBERTa model performs better on STS tasks compared to the base RoBERTa model and its variations.
- SimCSE models have better anisotropy.

3 Methodology

Both training and tests are done by using the authors' code written in python. There were a few bugs about utilizing GPUs but other than that, the base code is used without any modification. Google Colab is used for training.

The base code, python notebooks for training and test, and result outputs are publicly available on `https://github.com/gokg/SimCSE`

3.1 Model descriptions

There are 4 different models are trained: SimCSE-BERT-base (uncased) and SimCSE-RoBERTa-base, both as supervised and unsupervised. In tests, 4 trained and 6 pre-trained models are tested and compared with the results in the paper. All parameters are used in default values as stated in the paper to make comparison fair.

3.2 Datasets

Train datasets consist of two datasets. For unsupervised learning, 1 million sentences that consist of 19,800,591 words from English Wikipedia are sampled. For supervised learning, 275,600 sentence triplets are used, where each triplet has two similar and one different sentence. Supervised learning data is a merge of MNLI (3) and SNLI (2) datasets.

3.3 Experimental setup and code

As mentioned before, training and testing are done on Google Colab environment. Python notebooks are prepared for each training model and one for tests for all models. Simply running these notebooks are sufficient to replicate the trainings and tests.

3.4 Computational requirements

Training are done on Google Colab by using NVIDIA Tesla K80 GPUs. Due to limited access to GPUs, tests are done on local CPU with 16 core Intel Core i7-10700 CPU @ 2.90GHz. Training and test times are given in Table 1.

Table 1: Training and Testing Times

Task	Duration	Device
Training SimCSE-BERT base / SimCSE-RoBERTa base	∼4h 20m	GPU (Google Colab)
STS Testing for SimCSE-BERT base / SimCSE-RoBERTa base	∼40m	CPU (Local)
STS Testing for SimCSE-RoBERTa large	∼1h 20m	CPU (Local)
TL Testing for SimCSE-BERT base / SimCSE-RoBERTa base	∼3h 20m	CPU (Local)

Table 2: Published STS Results in the Paper

Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
Unsupervised models								
GloVe embeddings (avg.)	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
BERT base (first-last avg.)	39.70	59.38	49.67	66.03	66.19	53.87	62.06	56.70
BERT base -flow	58.40	67.10	60.85	75.16	71.22	68.66	64.47	66.55
BERT base -whitening	57.83	66.90	60.90	75.08	71.31	68.24	63.73	66.28
IS-BERT base	56.77	69.24	61.21	75.23	70.16	69.21	64.25	66.58
CT-BERT base	61.63	76.80	68.47	77.50	76.48	74.31	69.19	72.05
SimCSE-BERT base	**68.40**	**82.41**	**74.38**	**80.91**	**78.56**	**76.85**	**72.23**	**76.25**
RoBERTa base (first-last avg.)	40.88	58.74	49.07	65.63	61.48	58.55	61.63	56.57
RoBERTa base -whitening	46.99	63.24	57.23	71.36	68.99	61.36	62.91	61.73
DeCLUTR-RoBERTa base	52.41	75.19	65.52	77.12	78.63	72.41	**68.62**	69.99
SimCSE-RoBERTa base	**70.16**	**81.77**	**73.24**	**81.36**	**80.65**	**80.22**	68.56	**76.57**
SimCSE-RoBERTa large	**72.86**	**83.99**	**75.62**	**84.77**	**81.80**	**81.98**	**71.26**	**78.90**
Supervised models								
InferSent-GloVe	52.86	66.75	62.15	72.77	66.87	68.03	65.65	65.01
Universal Sentence Encoder	64.49	67.80	64.61	76.83	73.18	74.92	76.69	71.22
SBERT base	70.97	76.53	73.19	79.09	74.30	77.03	72.91	74.89
SBERT base -flow	69.78	77.27	74.35	82.01	77.46	79.12	76.21	76.60
SBERT base -whitening	69.65	77.57	74.66	82.27	78.39	79.52	76.91	77.00
CT-SBERT base	74.84	83.20	78.07	83.84	77.93	81.46	76.42	79.39
SimCSE-BERT base	**75.30**	**84.67**	**80.19**	**85.40**	**80.82**	**84.25**	**80.39**	**81.57**
SRoBERTa base	71.54	72.49	70.80	78.74	73.69	77.77	74.46	74.21
SRoBERTa base -whitening	70.46	77.07	74.46	81.64	76.43	79.49	76.65	76.60
SimCSE-RoBERTa base	**76.53**	**85.21**	**80.95**	**86.03**	**82.57**	**85.83**	**80.50**	**82.52**
SimCSE-RoBERTa large	**77.46**	**87.27**	**82.36**	**86.66**	**83.93**	**86.70**	**81.95**	**83.76**

4 Results

Semantic Text Similarity

This section contains STS test results.

The authors' results published in the paper are given in Table 2.

Since the best results are given by SimCSE models, we tested the pre-trained models to get the same results, which are shown in Table 3.

Then we trained the same models and compared with the pre-trained models. Results are shown in Table 4.

As a result, even though in some STS tests the difference went up to 5% difference, in the average results we could reproduce the achievements by 1-2% error margin.

Transfer Task

This section contains TT test results.

Table 3: Comparison of Pre-Trained Model STS Tests and Published STS Results

Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
Unsupervised models								
On-Paper SimCSE-BERT base	68.40	82.41	74.38	80.91	78.56	76.85	72.23	76.25
Self-Tested SimCSE-BERT base	68.40	82.41	74.38	80.91	78.56	76.85	72.23	76.25
On-Paper SimCSE-RoBERTa base	70.16	81.77	73.24	81.36	80.65	80.22	68.56	76.57
Self-Tested SimCSE-RoBERTa base	70.15	81.77	73.24	81.35	80.65	80.22	68.56	76.56
On-Paper SimCSE-RoBERTa large	72.86	83.99	75.62	84.77	81.80	81.98	71.26	78.90
Self-Tested SimCSE-RoBERTa large	72.86	84.00	75.62	84.77	81.80	81.99	71.26	78.90
Supervised models								
On-Paper SimCSE-BERT base	75.30	84.67	80.19	85.40	80.82	84.25	80.39	81.57
Self-Tested SimCSE-BERT base	76.53	85.20	80.95	86.03	82.56	85.83	80.50	82.51
On-Paper SimCSE-RoBERTa base	76.53	85.21	80.95	86.03	82.57	85.83	80.50	82.52
Self-Tested SimCSE-RoBERTa base	75.30	84.67	80.19	85.40	80.82	84.26	80.39	81.58
On-Paper SimCSE-RoBERTa large	77.46	87.27	82.36	86.66	83.93	86.70	81.95	83.76
Pre-Trained SimCSE-RoBERTa large	77.46	87.27	82.36	86.66	83.93	86.70	81.95	83.76

Table 4: Comparison of Self-Trained and Pre-Trained Models on STS Tasks

Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
Unsupervised models								
Pre-Trained SimCSE-BERT base	**68.40**	**82.41**	**74.38**	**80.91**	**78.56**	**76.85**	**72.23**	**76.25**
Self-Trained SimCSE-BERT base	65.25	81.29	71.26	79.18	76.34	75.60	69.30	74.03
Pre-Trained SimCSE-RoBERTa base	**70.15**	81.77	**73.24**	**81.35**	**80.65**	**80.22**	**68.56**	**76.56**
Self-Trained SimCSE-RoBERTa base	67.03	**82.20**	72.67	81.25	78.55	78.09	68.06	75.41
Supervised models								
Pre-Trained SimCSE-RoBERTa base	76.53	**85.20**	**80.95**	**86.03**	**82.56**	**85.83**	**80.50**	**82.51**
Self-Trained SimCSE-BERT base	**77.11**	79.22	78.07	85.30	81.50	82.13	78.88	80.32
Pre-Trained SimCSE-BERT base	75.30	**84.67**	**80.19**	**85.40**	80.82	**84.26**	**80.39**	**81.58**
Self-Trained SimCSE-RoBERTa base	**76.53**	77.04	77.89	83.55	**81.54**	82.38	75.96	79.27

The authors' results published in the paper are given in Table 5.

Since the best results are given by SimCSE models, we tested the pre-trained models to get the same results, which are shown in Table 6.

Then we trained the same models and compared with the pre-trained models. Results are shown in Table 7.

5 Discussion

Given that our trained models reproduced the results within only 2% error margin, we can say that published results can be reproduced in a confident margin. However, the fact that most of the time reproduced results were lower than the published ones had us thinking if the parameters published in the paper were not the most optimum ones, and a slightly updated set of parameters were used to produce the results.

5.1 Further tests

Our work did not include all the tests we aimed to reproduce. One of the main claims in the paper was that dropout mask as the only noise source is the most beneficial way, compared to the other noise techniques such as word deletion or sentence trimming. We also couldn't test dropout rate tests, which was another point made in the paper.

Table 5: Transfer Task Results Published in the Paper

Model	MR	CR	SUBJ	MPQA	SST	TREC	MRPC	Avg.
Unsupervised models								
Avg. BERT embeddings	78.66	86.25	94.37	88.66	84.40	92.80	69.54	84.94
BERT-[CLS] embedding	78.68	84.85	94.21	88.23	84.13	91.40	71.13	84.66
IS-BERT base	81.09	87.18	94.96	88.75	85.96	88.64	74.24	85.83
SimCSE-BERT base	81.18	86.46	94.45	88.88	85.50	89.80	74.43	85.81
w/ MLM	82.92	87.23	95.71	88.73	86.81	87.01	78.07	86.64
SimCSE-RoBERTa base	81.04	87.74	93.28	86.94	86.60	84.60	73.68	84.84
w/ MLM	83.37	87.76	95.05	87.16	89.02	90.80	75.13	86.90
SimCSE-RoBERTa large	82.74	87.87	93.66	88.22	88.58	92.00	69.68	86.11
w/ MLM	84.66	88.56	95.43	87.50	89.46	95.00	72.41	87.57
Supervised models								
SBERT base	83.64	89.43	94.39	89.86	88.96	89.60	76.00	87.41
SimCSE-BERT base	82.69	89.25	94.81	89.59	87.31	88.40	73.51	86.51
w/ MLM	82.68	88.88	94.52	89.82	88.41	87.60	76.12	86.86
SRoBERTa base	84.91	90.83	92.56	88.75	90.50	88.60	78.14	87.76
SimCSE-RoBERTa base	84.92	92.00	94.11	89.82	91.27	88.80	75.65	88.08
w/ MLM	85.08	91.76	94.02	89.72	92.31	91.20	76.52	88.66
SimCSE-RoBERTa large	88.12	92.37	95.11	90.49	92.75	91.80	76.64	89.61
w/ MLM	88.45	92.53	95.19	90.58	93.30	93.80	77.74	90.23

Table 6: Comparison of Pre-Trained Model TT Tests and Published TT Results

Model	MR	CR	SUBJ	MPQA	SST2	TREC	MRPC	Avg.
Unsupervised models								
On-Paper SimCSE-BERT base	**82.92**	**87.23**	**95.71**	88.73	**86.81**	87.01	**78.07**	**86.64**
Self-Tested SimCSE-BERT base	81.18	86.46	94.43	**88.87**	85.50	**89.80**	74.49	85.82
On-Paper SimCSE-RoBERTa base	**83.37**	**87.76**	**95.05**	**87.16**	**89.02**	**90.80**	**75.13**	**86.90**
Self-Tested SimCSE-RoBERTa base	81.04	87.74	93.28	86.94	86.60	84.60	73.68	84.84
Supervised models								
On-Paper SimCSE-BERT base	82.68	88.88	94.52	**89.82**	**88.41**	87.60	**76.12**	**86.86**
Self-Tested SimCSE-BERT base	**82.88**	**89.20**	**94.81**	89.67	87.31	**88.40**	73.51	86.54
On-Paper SimCSE-RoBERTa base	**85.08**	91.76	94.02	89.72	**92.31**	**91.20**	**76.52**	**88.66**
Self-Tested SimCSE-RoBERTa base	85.01	**92.00**	**94.05**	**89.84**	91.27	88.00	75.65	87.97

Table 7: Comparison of Self-Trained and Pre-Trained Models in TT Tasks

Model	MR	CR	SUBJ	MPQA	SST2	TREC	MRPC	Avg.
Unsupervised models								
Pre-Trained SimCSE-BERT base	**81.18**	**86.46**	**94.43**	**88.87**	**85.50**	**89.80**	**74.49**	**85.82**
Self-Trained SimCSE-BERT base	80.47	85.62	93.89	88.03	84.68	84.60	74.14	84.49
Pre-Trained SimCSE-RoBERTa base	**81.04**	**87.74**	**93.28**	**86.94**	**86.60**	**84.60**	**73.68**	**84.84**
Self-Trained SimCSE-RoBERTa base	79.45	84.85	91.79	85.75	84.90	83.00	74.14	83.41
Supervised models								
Pre-Trained SimCSE-BERT base	**82.88**	**89.20**	**94.81**	**89.67**	**87.31**	**88.40**	73.51	**86.54**
Self-Trained SimCSE-BERT base	82.42	88.48	94.15	89.54	86.93	87.00	**75.36**	86.27
Pre-Trained SimCSE-RoBERTa base	**85.01**	**92.00**	**94.05**	**89.84**	91.27	**88.00**	**75.65**	**87.97**
Self-Trained SimCSE-RoBERTa base	84.38	90.97	92.97	89.19	**91.32**	84.60	**75.65**	87.01

RoBERTa large models couldn't be trained due to limited computational resources. Only pre-trained large models are used in testing.

Lastly, not all models could be tested on Transfer Task tasks. These tasks take longer time than STS tasks and were not the focused tasks in the paper, therefore we didn't prioritised them in the resource usage.

References

[1] Gao, T., Yao, X. & Chen, D. SimCSE: Simple Contrastive Learning of Sentence Embeddings. *ArXiv Preprint ArXiv:2104.08821*. (2021)

[2] Bowman, S., Angeli, G., Potts, C. & Manning, C. A large annotated corpus for learning natural language inference. *ArXiv Preprint ArXiv:1508.05326*. (2015)

[3] Williams, A., Nangia, N. & Bowman, S. A broad-coverage challenge corpus for sentence understanding through inference. *ArXiv Preprint ArXiv:1704.05426*. (2017)