Poorya Piroozfar, Fatemeh Dokhanian
Iran University of Science and Technology, Tehran, Iran
Poorya_piroozfar@comp.iust.ac.ir, F_dokhanian@comp.iust.ac.ir

Scope of Reproducibility

One of the claims of the paper that we intend to tackle in this report is adjusting the fine-tuned embedding space for isotropy hurts its performance.
This report proves that the clustered structure of the embedding space changes during fine-tuning.
Another claim in the article that we have proved is that the number of elongated dominant directions significantly increases after fine-tuning.

Methodology

In this project, the author's code has been used to repeat the experiments, and in some cases, we have added our own code to it using the information in the article, for example, for the visualization section of locating the models in pre-training mode or fine-tuning the 3D. scatter plot is used. Google Collab is used to execute the code, and the tests in this article do not require any special fees. (Ram:12GB, Disk:80GB)

Results

Based on the experiments performed by us, which are the experiments in the original article, we were able to reproduce the claims and achievements of the article. Although the numbers obtained in the various parameters do not exactly correspond to the numbers given in the article, the same results can be deduced.

What was easy

There are many ways to implement this article, but we used the author code, and in this code, a number of methods, such as the cluster-based method, were clearly present in the code and were easy to implement.

What was difficult

Although there was access to the author code, in some cases we implemented part of the code ourselves, such as vector visualization representation using a scatter plot or the zero-mean method. We also did not have access to the article dataset at first, which we obtained using the Datasets library in python, which of course is not a difficult step. We also used the article information and code available for the cluster-based method to implement the cluster-ZM and Global-App methods.

Communication with original authors

In this work, we contacted the first author of the article and to resolve the ambiguity in some parts of the code and the article, we contacted her.

1 Introduction

In this research, we have tried to reproduce the results and achievements of this article, and this was done with the help of codes provided by the author, and in this article, we have tried to investigate the effects of fine-tuning operations on the Geometry of Embedding Space. This is done by measuring a criterion called isotropy as well as the Spearman correlation criterion.
The Spearman correlation criterion examines the performance of models in Semantic textual similarity task, which is characterized by numbers between 0 and 5.
In this work, transformer-based models called Bert and Roberta are used, which are used in two modes, pre-trained or fine-tuned, for the STS task.

2 Scope of reproducibility

The claims of the article, which it has tried to prove by experiments, are as follows:

Claim 1: Adjusting the fine-tuned embedding space for isotropy hurts its performance. For this purpose, we have fine-tuned the model based on Siamese architecture. The Spearman correlation and isotropy criteria were examined for 5 different pre-trained adjustments and the precise adjustment of BERT and RoBERTa.

Claim 2: The number of elongated dominant directions significantly increases after fine-tuning.
With an equal number of omitted directions, the positioned space with the adjusted model has less isotropy (isotropy) compared to the positioned space with the pre-trained model. This means that in order to have similar placement spaces in terms of isotropy, we have to remove more dominant directions from the placement model with the fine-tuned model.

Claim 3: The clustered structure of the embedding space changes during fine-tuning. To investigate this claim, we will use the visualization of the Pre-trained and Fine-tuned CWRs.

Claim 4: By eliminating the 100 and 700 least dominant directions from a total of 768 directions, we observe a slight drop in the performance compared to removing 12 top dominant directions.
The above claims are substantiated by the results reproduced in Section 4.

3 Methodology

In this project, the author's code has been used to repeat the experiments, and in some cases, we have added our own code to it using the information in the article, for example, for the visualization section of locating the models in pre-training mode or fine-tuning the 3D. acatter plot is used. Google Collab is used to execute the code, and the tests in this article do not require any special fees. (Ram:12GB, Disk:80GB, Run Type: GPU)

3.1 Model descriptions

In this work, we use BERT and ROBERTA models in 2 types of pre-trained and fine-tuned for embedding, and finally implement 4 methods on this embedding space.

The approaches used in this article are as follows:
Zero-mean. This method simply transfers all the embeddings to the center.
Clustering+ZM. Here, we first cluster embeddings and then separately make each cluster zero-mean

Global app. In this method, after making embeddings zero-mean, a few top dominant directions calculated using PCA are discarded.

Cluster-based app. Here, we first cluster embeddings and then make each cluster zero-mean individually. At the last step, dominant directions are calculated in each cluster and discarded.

3.2 Datasets

To analyze changes in a fine-tuned model, this study chooses Semantic Textual Similarity (STS) as the target task considering the STS-Benchmark dataset. This dataset includes 8628 pairs of sentences with similarity labels, of which 5749 pairs are train set sentences and 1500 pairs are dev data sentences, and 1379 pairs of sentences are test data.
STS is a semantic regression task in which the model needs to determine the similarity of two sentences in a paired sample. The label is a continuous range from 0 to 5. The evaluation criterion in this task is Pearson or Spearman correlation. In the main article, the performance of the methods is evaluated using the Spearman correlation.

3.3 Experimental setup and code

In our experiments for reproducing the main article's results, We analyzed the influence of fine-tuning on the embedding space of the base versions of BERT and RoBERTa. Both models have similar transformer based architectures, while RoBERTa has been trained with more training data and a slight difference in the optimization procedure. For the pre-trained setting, we use the models as feature extractors (the weights are frozen in this phase). Applying the mean-pooling method over the word embeddings, we obtain a sentence representation for every sample and consider the cosine similarity of the sentence representations as the textual similarity score. In the fine-tuning scenario, we fine-tune the models with a Siamese architecture that uses cosine similarity and the mean-pooling method for sentence representation. In our experiments, the batch size is set to 32, the learning rate is set to 7E-5, and the models are fine-tuned for 3 epochs. We set the number of clusters and discarded dominant directions in Global and Cluster-based approaches to 27 and 12, respectively, for both models.
Evaluation of Methods and Models in Experiments of this paper uses the Spearman correlation between estimated similarity scores and label scores (in the STS task).
Isotropy is a geometrical assessment of the distribution of data points in a feature space, which is ideally uniform. An embedding space is considered isotropic if the word embeddings are not biased towards a specific direction (feature). In other words, in isotropic space, word embeddings are uniformly distributed in space, leading to low correlation and near-zero cosine similarity for randomly sampled words.
Our code is available at https://github.com/fatemeh-dkh/NLP-article-challenge.

4 Results

Based on the experiments performed by us, which are the experiments in the original article, we were able to reproduce the claims and achievements of the article. Although the outcomes obtained in the various parameters do not exactly correspond to the outcomes given in the article, the same results can be deduced.

4.1 Result 1

For claim 1, based on the experiments performed, we have obtained the following results in table 1 that confirm claim 1.
Although the outcomes obtained do not exactly match the outcomes in the article, the results obtained in the article are also confirmed by our experiments.
Also, with the results obtained from the cluster-based and global methods, claim 2 can be proved.

	Baseline		Zero-mean		Clustering + ZM		Global		Cluster-based	
	Perf.	Isotropy	Perf.	Isotropy	Perf.	Isotropy	Perf.	Isotropy	Perf.	Isotropy
Pre-trained	61.40	1.7E-07	63.56	1.9E-05	59.41	0.30	73.55	0.41	75.37	0.74
Fine-tuned	79.40	3.9E-3	81.56	6.4E-3	81.32	0.09	79.55	0.18	68.63	0.64
Pre-trained	36.51	2.9E-6	35.43	7.9E-02	62.58	0.7	63.43	0.82	58.43	0.91
Fine-tuned	79.45	3.5E-4	79.03	5.9E-03	72.98	0.07	78.65	0.16	57.62	0.26

Table 1: Spearman correlation performance and isotropy for five different settings in the pre-trained and fine-tuned BERT and RoBERTa

4.2 Result 2

We have obtained the following results by testing claim 3.
As shown in Figure 1, in this project, the effect of fine-tuning on the structure of the BERT and RoBERTa embedding spaces was investigated.
The analyzes performed show that the significant performance usually obtained as a result of fine-tuning is not due to its increased isotropic in the embedding space. Like pre-trained examples, fine-tuned CWRs have oriented directions toward different dimensions across all layers, and the number of these directions increases with fine tuning.

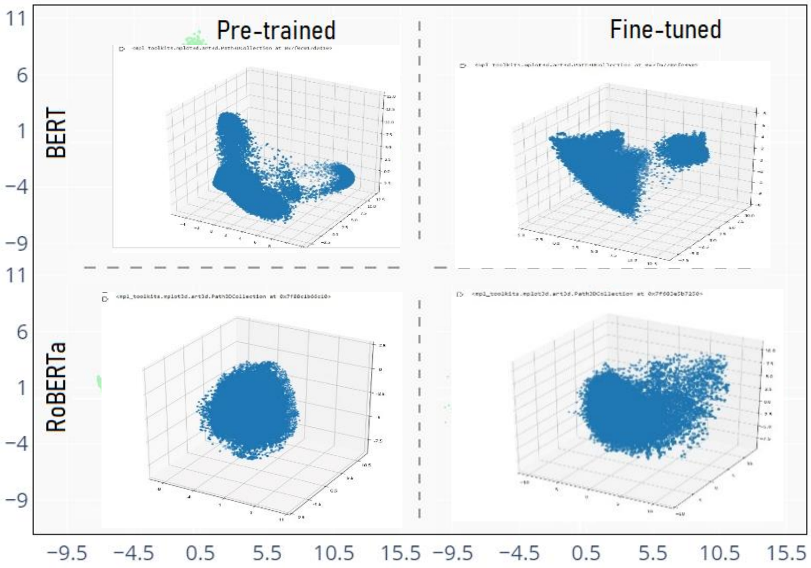

Figure 1: Illustration of pre-trained and fine-tuned CWRs

4.3 Result 3

Claim 4 can also be proved by comparing the results of Table 1 and Table 2. The results of table 2 do not exactly match the numbers reported in the main article, but they do support the main claim of the article about this experiment.

	Baseline		Global-app **100 least dir**		Global-app **700 least dir**		Cluster-based **100 least dir**		Cluster-based **700 least dir**	
	Perf.	Isotropy	Perf.	Isotropy	Perf.	Isotropy	Perf.	Isotropy	Perf.	Isotropy
BERT	79.40	3.9E-3	83.68	2.01E-3	80.85	2.35E-3	76.89	0.14	74.25	0.15
RoBERTa	79.45	3.5E-4	80.36	3.10E-4	76.25	1.32E-2	72.15	0.12	70.32	0.12

Table 2: Spearman correlation performance and isotropy after removing the least dominant directions in Global and Cluster-based approaches on STS dev set.

References

[1] Rajaee, S., & Pilehvar, M. T, “How Does Fine-tuning Affect the Geometry of Embedding Space: A Case Study on Isotropy,” In Findings of the Association for Computational Linguistics: EMNLP 2021 (pp. 3042-3049).

[2] S. Rajaee and M. T. Pilehvar, “A Cluster-based Approach for Improving Isotropy in Contextual Embedding Space,” Published in ACL/IJCNLP 2 June 2021 (pp. 575–584).