InterCSE: Sentences Representation with Interactional Framework of Supervised Sentence Representation

Anonymous EMNLP submission

Abstract

Contrastive learning has achieved remarkable results in sentence representation, but its semantic representation remains independent in the process of training and inference, and could not pay attention to the interactive information of sentence pairs. Therefore, this paper proposes to introduce a multi-task contrastive learning method, which not only focuses on the sorting of sentence pairs embedding similarity, but also increases the interaction information as a supplement to the sentence embedding representation. We evaluate the performance of InterCSE on several datasets, and experiments show that our model has a 0.40% improvement. Our code is available at https://github.com/2hip3ng/InterCSE.

1 Introduction

Sentence representation learning is a vital component of natural language processing tasks (Cer et al., 2017). The rapid development of sentence representation technology has made a wide range of downstream tasks more intelligent, especially information retrieval and text clustering.

Recently, the pre-trained language model has become the cornerstone of natural language processing technology, such as BERT(Devlin et al., 2019; Liu et al., 2019), GPT(Radford et al., 2018, 2019; Brown et al., 2020), ERNIE(Sun et al., 2019a,b, 2021), which greatly affects the development of various downstream tasks. Sentence representation could not get high performance directly with pre-trained language model becouse of anisotropic phenomenon (Gao et al., 2019), but contrastive learning play the role of a bridge. Many sentence representation approaches are transformed into pointwise classification tasks(Reimers and Gurevych, 2019), and there is a gap between the training optimization objectives and good sentence representation. Comparative learning applys the sorting method with InfoNCE(van den Oord et al., 2019)

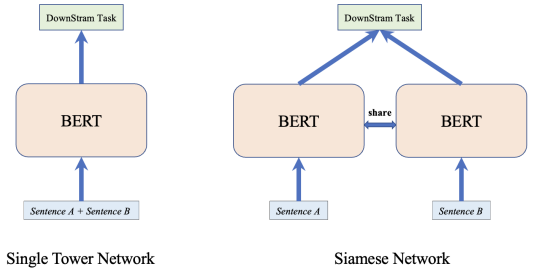

Figure 1: The architecture of single tower network(left) and siamese network(right).

contrastive loss function, which is more suitable for the optimization goal of sentence representation.

The sentence representation model(Gao et al., 2022; Yan et al., 2021) of contrastive learning combined with pre-trained model is a siamese network, which encodes all sentences independently. The structure of the siamese network makes it possible to quickly produce the vector representation of the sentences and calculate the similarity during training and prediction, so as to complete the retrieval or sentence clustering task in massive data. However, the structure of independent encoding makes the sentences pair lose the interactive information, which reduces the accuracy rate.

There are obvious differences between single tower network and siamese network. The network structure is depicted in Figure 1. The single tower network is like the original BERT(Devlin et al., 2019) model structure. After concating the sentence pairs, the semantic features of the sentence pairs are extracted and input to the downstream classification or regression network. The siamese network(Koch et al., 2015) has two encoders that encode each sentence individually. After encoding, it can perform similarity calculation or put into the downstream network using sentence embedding, such as Sentence-BERT(Reimers and Gurevych, 2019). During training, the two encoders share model parameters.

Network Type	Model	Spearman
Siamese Network	SBERT-base(Reimers and Gurevych, 2019)	84.67
	SBERT-large(Reimers and Gurevych, 2019)	84.45
	SimCSE-BERT-base(Gao et al., 2022)	84.25
	SimCSE-BERT-large(Gao et al., 2022)	-
	SRoBERTa-base(Reimers and Gurevych, 2019)	84.92
	SRoBERTa-large(Reimers and Gurevych, 2019)	85.02
	SimCSE-RoBERTa-base(Gao et al., 2022)	85.83
	SimCSE-RoBERTa-large(Gao et al., 2022)	86.70
Single Tower Network	BERT-base(Devlin et al., 2019)	85.8
	BERT-large(Devlin et al., 2019)	86.5
	RoBERTa-base(Ott et al., 2019)	87.2
	RoBERTa-large(Ott et al., 2019)	88.1

Table 1: Comparison of singel tower network and siamese network on STS-Benchmark(Cer et al., 2017) test set with spearman's correlation. SimCSE(Gao et al., 2022) models are trained on NLI datasets(Bowman et al., 2015), and the other are trained on STS-Benchmark train set. The performance of RoBERTa-base and RoBERTa-large in Single Tower Network are reproduce through fairseq(Ott et al., 2019).

The performance of the single-tower network and the siamese network on the STS-B test dataset is shown in Table 1. On the whole, the single tower network has a good improvement in spearman's correlation index compared with the siamese network. Especially on the **** model, the single tower network has such an improvement of ****. We consider that when the single tower network encodes a sentence pair, the sentence is not an independent individual. It will refer to its counterparts for encoding and use the attention mechanism to extract features, which makes the single tower network in the sentence pair similarity task has a higher correlation coefficient. Recently, most sentence representation tasks use siamese networks. Although it could bring computational advantages on massive data, it inevitably reduces the accuracy of correlation.

In order to take full advantage of the accuracy advantages of single tower networks and the inference speed advantages of Siamese networks, this paper proposes a multi-task contrastive learning method. During the training process, on the basis of SimCSE, we added a single tower network as a supplement to form a framework for multi-task learning. This method can fully obtain the interaction information between sentence pairs during the process of training sentence representation.

Our contributions can be summarized as follows:

1. We propose a framework for multi-task contrastive learning for sentence representation tasks.
2. We design many different loss functions for the proposed framework.
3. Our approach achieves new state-of-the-art performance on STS tasks.

2 Related Work

2.1 Large Language Model

Recently, transformer(Vaswani et al., 2017) structure shines in the field of deep learning and lays the foundation for large models. The attention mechanism is good at capturing the semantic relationship between sequences. Using transformer's encoder, many language models have been born, such as GPT(Radford et al., 2018), BERT(Devlin et al., 2019), RoBERTa(Liu et al., 2019), XLNet(Yang et al., 2019), Ernie(Sun et al., 2019a), etc. These models mainly have some differences in masking mechanism, pre-training data, and pre-training methods. Among them, the BERT model mainly masks a little words in the sentence and predicts the origin words as a unsupervised pre-training task. After obtaining the pre-trained model, NLP downstream tasks only need to complete fine-tuning on the model to achieve high performances and set new state-of-the-art results, including sentence classification, sequence tagging, question answering, machine reading and comprehension and sentence-pair classification or regression.

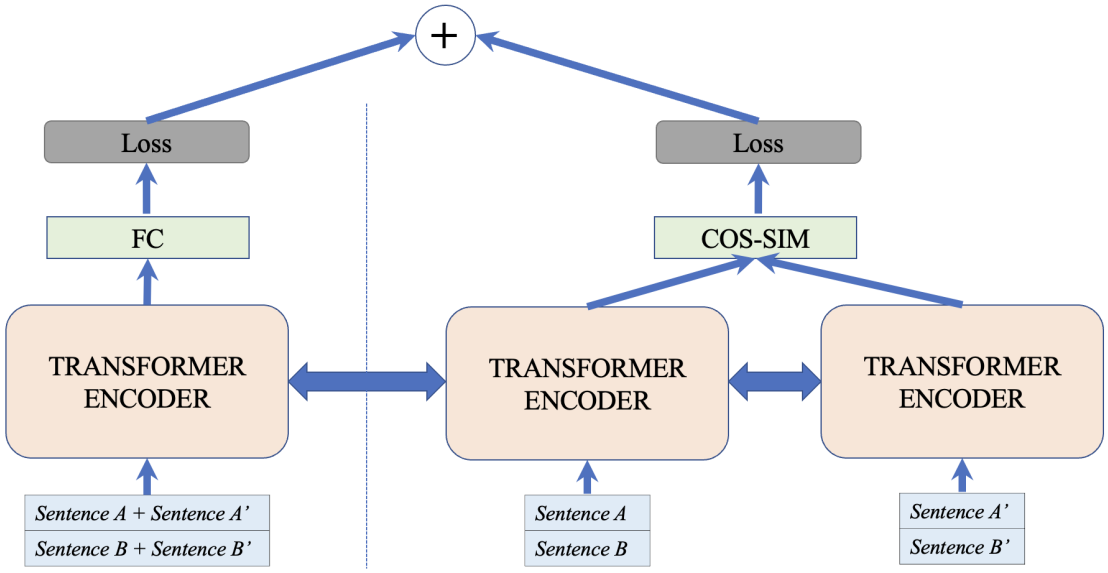

Figure 2: The architecture of InterCSE.

2.2 Contrastive Learning and Sentence Representation

Contrastive learning was first proposed in the field of computer vision. It mainly wants to optimize the similarity of sample pairs in a representation space, and make similar sample pairs gather and dissimilar sample pairs stay away. The core idea is that in a batch of sample pairs, there is only one pair of samples that are positive samples, and the others are negative samples. During the training process, after the sample representation is obtained using Siamese network coding, the sample similarity is calculated to maximize the similarity of the positive samples. Equation 1 can be used to represent the loss function, which is also referred to as **. This optimization scenario is very suitable for unsupervised scenarios. Usually, positive samples can be obtained through simple data enhancement, and a large number of negative samples can be obtained through negative sampling.

In terms of text representation, the idea of contrastive learning continues to be used, leading to many research results. How about ConBert, how about SimCSE, how about ESimCSE

2.3 Multi-Task Learning

Multi-Task Learning (MTL) is an important research topic in Machine Learning, which aims to learn multiple tasks simultaneously.In MTL, multiple tasks are divided into multiple learning units, and a learner can learn multiple tasks by making multiple small models. Compared with traditional multi-task learning, MTL can help improve the learning efficiency, reduce the resource consumption, and reduce the computational complexity.

One of the important research directions in MTL is to develop efficient algorithms and theoretical models. In recent years, many works have been carried out in this direction, including deep neural networks, attention mechanism, joint attention, graph attention, and so on.These algorithms and theoretical models can help MTL learn multiple tasks with better performance.

3 Approach

In this section, we present InterCSE for sentence representation task. First, we introduce the overall framework of the model. Then, we introduce the design idea of loss function for multi-task learning.

3.1 Framework

Our approach is mainly inspired by GraphSage(Hamilton et al., 2018). Transformer sage layer aggregates the embedding of neighbor nodes, and realizes an information interaction between neighbors, which enhances the representation ability of nodes. The main idea of our method is to introduce information interaction when encoding sentence, so a single tower network is added on the basis of SimCSE(Gao et al., 2022). During the training process, the single tower network completes the interactive feature extraction of sentence pairs.

As show in Figure 2, there are two major components in our framework: interactive network and independent network.

Interactive Network The goal of the interaction network is to obtain non-independent text semantic information through the attention interaction of words between text pairs. Non-independent text semantic information is an important correlation signal, which can keenly perceive the semantics of text to details. The interactive network consists of an encoder and a classifier. The encoder is a BERT-type sub-network, and the input is the splicing result of text A and text B. At the same time, the overall semantic information after encoding, that is, the result of CLS is input into the classifier, which is usually a forward network to evaluate the similarity or similar classification of text pairs.

Independent Network The goal of the independent network is to quickly produce semantic representations of independent texts without relying on other texts. The independent network is composed of an encoder and a similar network calculation. The encoder is a twin network whose sub-network is a BERT type model. The similar network is generally cosine similarity calculation or dot product calculation. During the training process, it uses the twin-tower network to encode the text pairs independently, and calculates the similarity between the encoded Embeddings.

3.2 Loss Expression

The model we propose is a multi-task model, including two modules of interactive network and independent network. The modeling objectives of the two modules are different, and the corresponding loss functions are also different. The following describes the loss functions corresponding to the two modules in detail.

For the interaction module, its input is a text pair with known classification labels or ranking labels. Therefore, the loss functions of the interactive network are mainly: classification loss function and sorting loss function. Among them, we use the cross-entropy loss function for the classification loss function, and logloss for the sorting loss function.

For the independent encoding module, its input is a batch of data, one pair of texts is a positive sample, and the rest of the text is a negative sample, that is, In-batch negative sampling. Its loss function can be expressed as:

Multi-task learning needs to integrate different loss functions. We propose the following two methods to regularize three different loss functions:

The first method simply adds up the various loss functions with different weights. The second method adds a penalty term to the sub-loss functions respectively.

4 Experiments

Our approach is mainly proposed for supervised tasks, and we conducted multiple experiments on Semantic Textual Similarity (STS) task to verify the effectiveness of this approach.

4.1 Setups

Datasets Following previous works(Reimers and Gurevych, 2019; Gao et al., 2022; Yan et al., 2021), we evaluate our approach on 7 STS tasks: STS 2012-2016(Agirre et al., 2012, 2013, 2014, 2015, 2016), STS Benchmark(Cer et al., 2017) and SICK-Relatedness(Marelli et al., 2014). Each sample in these datasets contains a pair of sentences as well as a gold score between 0 and 5 to indicate their semantic similarity. The higher the score, the higher the similarity of sentences pair.

Since the label (real number) of the STS dataset is not suitable for the training process of the contrastive learning model, we introduce the SNLI(Bowman et al., 2015) and MNLI(Williams et al., 2018) datasets as supervisory signals to train our model.

Baselines To show the effectiveness of our approach on supervised sentence representation, we select many state-of-the-art method as comparison recently, including InferSent(Conneau et al., 2017), Universal Sentence Encoder(Cer et al., 2018), Sentence-BERT(Reimers and Gurevych, 2019), ConSERT(Yan et al., 2021), SimCSE(Gao et al., 2022).

Evaluation When evaluating the trained model, we first obtain the representation of sentences by CLS token, then we report the spearman correlation between the cosine similarity scores of sentence representations and the human-annotated gold scores. When calculating spearman correlation, we merge all sentences together (even if some STS datasets have multiple splits) and calculate spearman correlation for only once.

4.2 Training Details

To Be Continued.

4.3 Main Results

To Be Continued.

5 Analysis

To Be Continued.

6 Conclusion

To Be Continued.

Limitations

1. Failed to train on sts-b training set.
2. To Be Continued.

Acknowledgements

To Be Continued

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. SemEval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91, Dublin, Ireland. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, Montréal, Canada. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43, Atlanta, Georgia, USA. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2019. Representation degeneration problem in training natural language generation models.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2022. Simcse: Simple contrastive learning of sentence embeddings.

William L. Hamilton, Rex Ying, and Jure Leskovec. 2018. Inductive representation learning on large graphs.

Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. 2015. Siamese neural networks for one-shot image recognition.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, Weixin Liu, Zhihua Wu, Weibao Gong, Jianzhong Liang, Zhizhou Shang, Peng Sun, Wei Liu, Xuan Ouyang, Dianhai Yu, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019a. Ernie: Enhanced representation through knowledge integration.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2019b. Ernie 2.0: A continual pre-training framework for language understanding.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. Consert: A contrastive framework for self-supervised sentence representation transfer.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. Cite arxiv:1906.08237Comment: Pretrained models and code are available at https://github.com/zihangdai/xlnet.