Contrastive Learning of Sentence Representations

Hefei Qiu Wei Ding Ping Chen
University of Massachusetts Boston, Boston, MA
{hefei.qiu001, wei.ding, ping.chen}@umb.edu

Abstract

Learning sentence representations which capture rich semantic meanings has been crucial for many NLP tasks. Pre-trained language models such as BERT have achieved great success in NLP, but sentence embeddings extracted directly from these models do not perform well without fine-tuning. We propose Contrastive Learning of Sentence Representations (CLSR), a novel approach which applies contrastive learning to learn universal sentence representations on top of pre-trained language models. CLSR utilizes semantic similarity of two sentences to construct positive instance for contrastive learning. Semantic information that has been captured by the pre-trained models is kept by getting sentence embeddings from these models with proper pooling strategy. An encoder followed by a linear projection takes these embeddings as inputs and is trained under a contrastive objective. To evaluate the performance of CLSR, we run experiments on a range of pre-trained language models and their variants on a series of Semantic Contextual Similarity tasks. Results show that CLSR gains significant performance improvements over existing SOTA language models.

1 Introduction

Learning sentence representations that can encode semantic information is crucial for many Natural Language Processing (NLP) tasks such as question answering, summarization, machine translation. Many attempts have been made to learn general purpose sentence embeddings (Le and Mikolov, 2014; Kiros et al., 2015; Hill et al., 2016; Conneau et al., 2017; Arora et al., 2017; Logeswaran and Lee, 2018; Cer et al., 2018; Subramanian et al., 2018; Pagliardini et al., 2018). Since transformer-based pre-trained language models such as BERT are introduced and achieve the state-of-the-art results in many NLP tasks, several methods have been proposed to generate sentence embeddings with some pooling strategy such as mean, max from word level embeddings and fine tune these models on downstream tasks (Reimers and Gurevych, 2019) or train to calibrate these models for isotropic embeddings (Li et al., 2020; Su et al., 2021).

Inspired by the recent development of contrastive learning in learning visual representations (Chen et al., 2020), we design a contrastive learning based architecture to learn high-quality semantic sentence representations and show this approach can significantly improve the sentence representations. We integrate both pre-trained language models and contrastive learning and call our architecture Contrastive Learning of Sentence Representations (CLSR). Different from previous methods of using data augmentation to construct positive pairs in contrastive learning (Chen et al., 2020), we use semantic similarity or entailment relation of two sentences to build positive pairs. By sending two similar sentences into a pre-trained language model and then generate vector representations by pooling, CLSR is able to keep the rich information captured by these pre-trained models. A contrastive objective is used to train an encoder followed by a linear projection to further learn the embeddings in an unsupervised way. CLSR is model-agnostic. The initial sentence embeddings it takes as inputs can come from any pre-trained model.

2 Related Work

Learning sentence embeddings has attracted a lot of interest. Paragraph Vector (Le and Mikolov, 2014) and Skip-Thought (Kiros et al., 2015) learns generic, distributed sentence representations in an unsupervised fashion, one by proposing two log-bilinear models and the other one by training an encoder-decoder model to reconstruct the surrounding sentences of an encoded passage. Hill

Proceedings of the 18th International Conference on Natural Language Processing, pages 277–283
Silchar, India. December 16 - 19, 2021.

et al. (2016) proposed Sequential Denoising Autoencoders and FastSent to learn sentence representations from unlabelled data. InferSent (Conneau et al., 2017) performs the learning in a supervised way by training a Siamese BiLSTM network on Stanford Natural Language Inference (SNLI) dataset (Bowman et al., 2015). Universal Sentence Encoder (Cer et al., 2018), with two variants, combines both by training with transfer learning on unsupervised data and being augmented on supervised data from the SNLI corpus (Bowman et al., 2015).

Recently there have been several attempts to improve sentence embeddings from pooled outputs of pre-trained language model such as BERT (Devlin et al., 2019). Sentence-BERT (Reimers and Gurevych, 2019) trains a Siamese network to fine-tune BERT and its variants. BERT-flow (Li et al., 2020) learns more smooth and isotropic embeddings by applying normalizing flow (Kumar et al., 2020) to convert BERT sentence embedding distribution into a Gaussian distribution. SBERT-WK (Wang and Kuo, 2020) improves Sentence-BERT by incorporating the pattern of layer-wised word representations in subspace. Su et al. (2021) applied whitening technique to enhance the isotropy of sentence representations.

Since contrastive learning has achieved great success in unsupervised representation learning in computer vision, it also has gained much interest in NLP including sentence representation learning. The following are some recent or concurrent work, many of which have explored different data augmentation strategies. IS-BERT (Zhang et al., 2020) learns through maximizing the mutual information between the global sentence representation and its local token representation. CERT (Fang and Xie, 2020) augments sentences using back-translation (Edunov et al., 2018). DeCLUTR (Giorgi et al., 2020) applies a contrastive objective on textual segments sampled from nearby in the same document. CLEAR (Wu et al., 2020) uses multi sentence-level augmentation to construct positive pairs. Carlsson et al. (2021) proposes Contrast Tension which counters the task biases in pre-trained language models by contrasting the noise between the output from two independent models. ConSERT (Yan et al., 2021) explores several ways to augment data such as adversarial attack, token shuffling etc. SimCSE (Gao et al., 2021) constructs positive instances by taking different outputs of the same sentence from the same pre-trained language model using dropout. Besides treating the task as unsupervised learning, although some of the above work such as Yan et al. (2021), Gao et al. (2021) also explore it as supervised learning, our method is mainly to show, by simply using sentence pairs with high similarity or entailment relation in existing labeled corpus to construct positive instances, contrastive learning can still further significantly improve the quality of sentence embeddings on top of any pre-trained language model.

3 Model

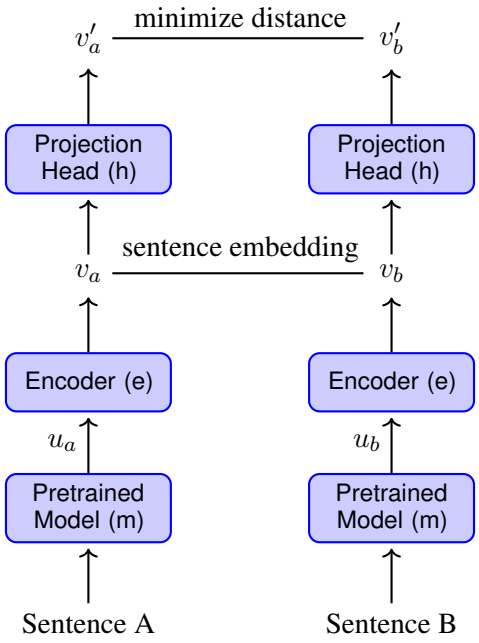

Figure 1: Contrastive learning of sentence representations. Two semantically similar sentences A and B (a positive pair) are sent to the same pre-trained model m (e.g., BERT). Each sentence embedding from m goes to an encoder e and a projection function h. A contrastive loss is applied to minimize the distance of two sentence embeddings. Output from e is used as the sentence representation for downstream tasks.

Contrastive learning has been a promising approach in self-supervised learning. It learns generic representations by contrasting positive pairs against negative pairs. SimCLR (Chen et al., 2020) is a simple framework for contrastive self-supervised learning of visual representations without using specialized architectures or a memory bank.

CLSR adopts contrastive learning framework SimCLR (Chen et al., 2020) to learn sentence semantic representations as shown in Figure 1. CLSR consists of:

- A pre-trained language model or pre-trained sentence encoder takes a raw sentence as input and output the sentence embedding, pooling may be applied. For example, pre-trained BERT with average pooling could be used to generate a vector representation for a sentence. Two sentences with high similarity are sent into the pre-trained model respectively to generate two sentence embeddings. They are considered as a positive pair. One is treated as a positive instance of the other. This step is different from SimCLR. We do not apply data augmentation technique to construct positive instance due to natural languages being highly discrete semantically. Instead, we use the property of semantic similarity.

- A neural network based encoder e further encodes sentence pairs into vectors respectively. This encoder could be of any structure. Since a well-pre-trained language model has been applied in the first step, in our setting, a simple Multi-Layer Perceptron (MLP) with 1 hidden layer and ReLu nonlinearity on the output is used to further encode information learned through contrastive learning without fine-tuning. The representation from this encoder is used for downstream tasks.

- A linear projection head h is applied to map sentence embeddings to a new representation space by training with contrastive objective. We follow the design in (Chen et al., 2020) which shows that a projection head can improve performance on downstream tasks.

- A contrastive loss function is defined as following: for a given set of sentences $\{s_m\}$ that contain positive sentence pair s_a and s_b with high semantic similarity, the contrastive learning process is to find s_b in $\{s_m\}_{b \neq a}$.

After randomly sampling m sentence pairs into a mini batch, this batch contains 2m sentences. Each sentence pair is treated as a positive pair, and the rest of sentences in the batch are treated as in-batch negatives (Chen et al., 2020; Henderson et al., 2017). We hypothesize that the probability of having one or more sentences in the negative samples that are highly semantically similar with that in the positive pair is very low and thus is ignored. Then the loss function of a positive sentence pair (s_a, s_b) is defined as:

$$\ell_{s_a,s_b} = -log \frac{e^{sim(v_a,v_b)/\tau}}{\sum_{i=1}^{2m} I_{(a,i)} e^{sim(v_a,v_i)/\tau}}, \quad (1)$$

where τ is the temperature hyper-parameter, sim(v_a, v_b) is a function measuring similarity between two given sentence vectors. We use cosine similarity for measurement. I is an indicator function to determine if a sentence s_i is included as a negative sample or not. If i $\neq$ a, it returns 1; otherwise, 0. The final loss is computed across all the positive sentence pairs in a mini batch. Positive pairs (s_a, s_b) and (s_b, s_a) are both included.

Base Model		CLSR-STSB	CLSR-NLI
BERT-base	59.32	64.94	**76.74**
BERT-large	57.77	63.84	**78.75**
SBERT-base	77.12	80.15	**81.93**
SBERT-large	79.19	80.19	**83.76**

Table 1: Spearman correlations on STS-B development set. The 1^{st} column includes 4 base models and their performance on STS-B. The 2^{nd} and 3^{rd} columns include the performance of CLSR built on each base model and trained on STS-B or NLI data respectively.

Batch Size	64	128	256	512	1024
STS-B	64.93	67.38	69.78	76.74	78.55

Table 2: Spearman correlations on STS-Benchmark development set to show the effect of batch size. CLSR is built on BERT-base-uncased and trained on STS-B.

4 Experiments

4.1 Experiment Settings

Datasets: We use two types of training data. One is Semantic Textual Similarity (STS) in which STS-Benchmark is chosen. It comprises 8,628 sentence pairs with similarity score 0-5. We pre-process the data by keeping sentence pairs with scores higher than 4 which indicate high similarity. This gives us totally 1,406 pairs. The other one is Natural Language Inference (NLI) data. We follow Sentence-BERT (Reimers and Gurevych, 2019) to concatenate two NLI datasets: SNLI (Bowman et al., 2015) and Multi-Genre NLI (Williams et al., 2018). SNLI dataset contains 570k English sentence pairs labeled with entailment, contradiction, and neutral. While Multi-Genre NLI is a collection of 433k sentence pairs annotated with the same three labels.We pre-process them by only selecting sentence pairs with label entailment. This leads to total 314,315 samples used in our training.

Model	STS12	STS13	STS14	STS15	STS16	STS-B	Avg.
BERT-base	45.61	56.57	57.15	62.94	64.74	64.48	58.58
CLSR-BERT-base	**59.83**	**66.16**	**63.80**	**70.11**	**69.71**	**70.03**	**66.61**
BERT-large	46.98	52.88	49.56	56.63	61.64	65.37	57.51
CLSR-BERT-large	**60.02**	**63.19**	**62.74**	**68.81**	**71.78**	**73.53**	**66.68**
SBERT-base	66.35	73.76	**73.88**	77.33	73.62	**73.63**	73.10
CLSR-SBERT-base	**66.38**	**74.96**	73.81	**77.93**	**75.22**	70.35	**73.11**
SBERT-large	68.79	75.71	**75.12**	**80.29**	75.91	75.35	75.20
CLSR-SBERT-large	**69.49**	**76.74**	74.64	78.90	**77.84**	**76.84**	**75.74**

Table 3: Comparison of CLSR models and their corresponding base pre-trained models on the series of STS tasks. Spearman correlations multiplied by 100 are reported. SBERT-base and SBERT-large refer to Sentence BERT built on BERT base or large and trained using NLI datasets with mean pooling strategy.

4 SOTA models for comparison: For an accurate assessment, BERT-base, BERT-large, SBERT-base, SBERT-large are selected to compare with CLSR. BERT models are selected due to their good performance and popularity in NLP. SBERT models are popular pre-trained sentence embedding models and represent the best SOTA performance.

6 STS tasks for evaluation: Since CLSR can take sentence embeddings from any base models such as BERT as input, we evaluate its effectiveness by comparing CLSR and its base models on the same task. For example, when we perform evaluation on STS-B, if CLSR is trained by taking embeddings from pre-trained BERT base model with mean pooling over its last layer, we run both BERT and CLSR models on STS-B. As Pearson correlation is shown to be not suitable for STS (Reimers et al., 2016), we report the results of Spearman correlation between the cosine similarity of a sentence pair and the ground truth label. We evaluate our model on 6 Semantic Textural Similarity tasks that include STS12-STS16 (Agirre et al., 2012, 2013, 2014, 2015; Artetxe et al., 2016), the STS-Benchmark (Cer et al., 2017).

Model setting and hyper-parameters: To fully assess the computation efficiency of our approach, we use a simple MLP with only 1 hidden layer and a 768-dimensional output layer as the encoder network, and a linear projection head to project the presentation to a latent space but with the same dimensions. We train at batch size 512 for 2000 epochs. Learning rate is 0.5 with the decay rate of 0.0001. Temperature is 0.1. We adopt linear warmup in the first 10 epochs and decay learning rate with cosine decay schedule without restarts (Chen et al., 2020; Loshchilov and Hutter, 2016).

4.2 Training Set Construction

To construct positive instances for contrastive learning, sentence pairs with high similarity in STS tasks can be naturally used. Sentences with entailment relation in NLI tasks can be an alternative option. In order to decide which training dataset performs better, we run an experiment on selection of training data. CLSR are trained with the 4 base models on STS-B and NLI datasets respectively. Following convention, Spearman correlations are reported on the STS-B development set. As shown in Table 1, compared with base pre-trained models, CLSR built on those models achieve significant improvements overall. Models trained on NLI dataset perform much better than those trained on STS-B with 15 points increment over BERT-base. Based on this result, we will report results only on NLI dataset due to space limitation.

Pooling Strategy	STS-B
CLS	61.34
mean	76.74

Table 4: Comparison on STS Benchmark development set to show the effect of different pooling strategy. The CLSR is built on BERT-base-uncased.

4.3 Results on STS Tasks

We evaluate CLSR framework on a series of STS tasks. We run a CLSR model and its base pre-trained model on the tasks respectively. Spearman correlations are reported in Table 3.

Compared with pre-trained BERT base and large models, the corresponding CLSR models increase the performance on all STS tasks by large margins. Compared with SOTA SBERT models, CLSR also shows solid improvement. It's reasonable to infer that such an improvement could be more significant with more training data, as we only train the CLSR model using roughly 1/3 of the NLI data, while SBERT fine-tunes the BERT models using all the data from the same dataset. The pre-trained BERT models are not fine-tuned and only a simple 2-layer MLP is designed to further encode the sentence

embedding. Surprisingly this simple approach can still slightly improve performance on several tasks compared with SBERT. This further validates the effectiveness of contrastive learning approach.

4.4 Ablation Study

Effect of batch size. Effect of batch size is shown in Table 2. The performance on STS-B development set shows that larger batch size brings better performance. This is consistent with the previous finding that contrastive learning benefits from large batch size (Chen et al., 2020). Since there are 2(n-1) negative instances in a mini-batch with size n, the change of batch size affects the number of negative instances more than that of positive instances. Thus it can be further inferred that, contrastive learning in the proposed framework learns more from larger number of negative instances.

Effect of pooling strategy. Previous work has shown the effect of pooling strategy (Xiao, 2018; Reimers and Gurevych, 2019). More specifically, taking the average of all the output word embeddings outperforms usage of the CLS token embedding as sentence embedding. This is also confirmed in our model as shown in table 4. By taking the mean of all the work embeddings from the last layer in BERT-base as input for CLSR, its performance on STS-Benchmark task increases as much as 15 points over the CLS token embedding.

5 Conclusion and Future Work

This paper presents a novel approach of applying contrastive learning on pre-trained language models to learn generic sentence representations. The evaluation on a series of STS tasks shows that our approach outperforms the pre-trained SOTA language models significantly. How to construct multiple positive instances and further integrate the idea of contrastive learning will be explored in future.

Acknowledgments

This work was supported by a NSF award IIS-1914489.

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, 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.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294, Austin, Texas. 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. *CoRR*, abs/1508.05326.

Fredrik Carlsson, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Hellqvist, and Magnus Sahlgren. 2021. Semantic re-tuning with contrastive tension. In *International Conference on Learning Representations*.

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.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

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. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. *CoRR*, abs/1808.09381.

Hongchao Fang and Pengtao Xie. 2020. CERT: contrastive self-supervised learning for language understanding. *CoRR*, abs/2005.12766.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings.

John M. Giorgi, Osvald Nitski, Gary D. Bader, and Bo Wang. 2020. Declutr: Deep contrastive learning for unsupervised textual representations. *CoRR*, abs/2006.03659.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun hsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1367–1377, San Diego, California. Association for Computational Linguistics.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *NIPS*, pages 3294–3302.

Manoj Kumar, Mohammad Babaeizadeh, Dumitru Erhan, Chelsea Finn, Sergey Levine, Laurent Dinh, and Durk Kingma. 2020. Videoflow: A conditional flow-based model for stochastic video generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1188–1196, Bejing, China. PMLR.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130, Online. Association for Computational Linguistics.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*.

Ilya Loshchilov and Frank Hutter. 2016. SGDR: stochastic gradient descent with restarts. *CoRR*, abs/1608.03983.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2018. Unsupervised learning of sentence embeddings using compositional n-gram features. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 528–540, New Orleans, Louisiana. Association for Computational Linguistics.

Nils Reimers, Philip Beyer, and Iryna Gurevych. 2016. Task-oriented intrinsic evaluation of semantic textual similarity. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 87–96, Osaka, Japan. The COLING 2016 Organizing Committee.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *CoRR*, abs/2103.15316.

Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J Pal. 2018. Learning general purpose distributed sentence representations via large scale multi-task learning.

Bin Wang and C.-C. Jay Kuo. 2020. SBERT-WK: A sentence embedding method by dissecting bert-based word models. *CoRR*, abs/2002.06652.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. Clear: Contrastive learning for sentence representation.

Han Xiao. 2018. bert-as-service. https://github.com/hanxiao/bert-as-service.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5065–5075, Online. Association for Computational Linguistics.

Yan Zhang, Ruidan He, Zuozhu Liu, Kwan Hui Lim, and Lidong Bing. 2020. An unsupervised sentence embedding method by mutual information maximization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1601–1610, Online. Association for Computational Linguistics.