Contrastive Learning with Prompt-derived Virtual Semantic Prototypes for Unsupervised Sentence Embedding

Jiali Zeng[1], **Yongjing Yin**[2]*, **Yufan Jiang**[1], **Shuangzhi Wu**[1], **Yunbo Cao**[1]
[1]Tencent Cloud Xiaowei, Beijing, China
[2]Westlake University, Zhejiang, China
{lemonzeng,garyyfjiang,frostwu,ethanlli}@tencent.com
yinyongjing@westlake.edu.cn

Abstract

Contrastive learning has become a new paradigm for unsupervised sentence embeddings. Previous studies focus on instance-wise contrastive learning, attempting to construct positive pairs with textual data augmentation. In this paper, we propose a novel **Con**trastive learning method with **P**rompt-derived **V**irtual semantic **P**rototypes (ConPVP). Specifically, with the help of prompts, we construct virtual semantic prototypes to each instance, and derive negative prototypes by using the negative form of the prompts. Using a prototypical contrastive loss, we enforce the anchor sentence embedding to be close to its corresponding semantic prototypes, and far apart from the negative prototypes as well as the prototypes of other sentences. Extensive experimental results on semantic textual similarity, transfer, and clustering tasks demonstrate the effectiveness of our proposed model compared to strong baselines. Code is available at https://github.com/lemon0830/promptCSE.

1 Introduction

High-quality sentence embeddings can boost the performance of pre-trained language models (PLMs) on many downstream tasks (Kiros et al., 2015; Logeswaran and Lee, 2018; Reimers and Gurevych, 2019a). Recent research focuses on learning sentence embeddings in an unsupervised manner due to lack of large scale labeled data (Hill et al., 2016; Pagliardini et al., 2018; Wang et al., 2021b). Among these methods, contrastive learning has been extensively explored and achieved remarkable success (Gao et al., 2021b; Wu et al., 2021; Yan et al., 2021; Giorgi et al., 2021). Specifically, most of them construct a positive pair by operating various textual data augmentation methods, while regard two independent sentences sampled uniformly from the training data as a negative pair. In spite of effectiveness in easing the anisotropy problem, such instance-wise optimization leads to a locally smooth embedding space, and ignores semantic relevance to some extent (Li et al., 2020b). Moreover, due to the discrete nature of language, data augmentation can change sentence semantics significantly, and thus a positive sample are possibly turned into a negative one (Wang et al., 2021a).

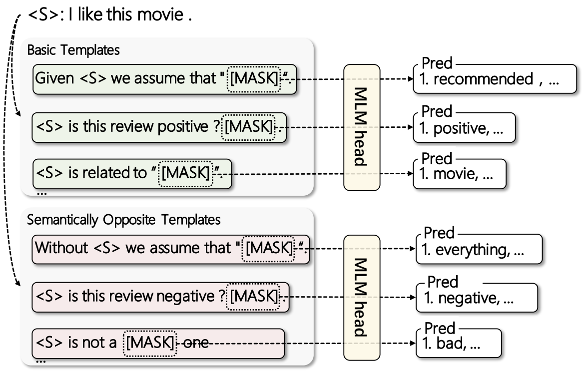

Figure 1: Paradigm of prompt learning.

To alleviate the issues, we introduce the idea of prototypical contrastive learning to unsupervised sentence embeddings learning, which is proven effective to learn structural visual embedding space (Li et al., 2020b; Caron et al., 2020). The motivation lies in that when encoding sentences into the embedding space, the sentences with similar semantics cluster together around the corresponding prototype. Nevertheless, the acquisition of prototypes is inefficient if we directly apply the clustering algorithms used in (Li et al., 2020b; Caron et al., 2020), due to the requirement of extra forward pass over the training set or weak correlation with semantics (Caron et al., 2020). This makes us wonder **whether there exists a dedicated method of mining the semantic prototypes for sentence embeddings especially based on PLMs?**

To answer this question, we attempt to think from the perspective of prompt learning (Brown et al., 2020). Intuitively, on sentence-level NLP tasks such as classification, the neural models en-

*Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 7071 - 7082
December 7-11, 2022 ©2022 Association for Computational Linguistics

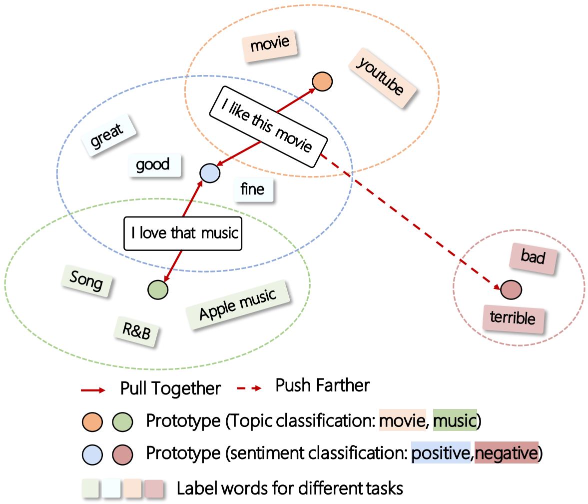

Figure 2: Illustration of contrastive learning with prompt-derived virtual semantic prototypes.

code and map each input to a corresponding semantic prototype in the embedding space. For example, the sentiment analysis models divide instances into semantic prototypes related to sentiment polarity. In addition, the sentence-level tasks can be solved by providing task-specific prompts to PLMs as a condition even without any fine-tuning (Brown et al., 2020; Sanh et al., 2021; Wei et al., 2021). As illustrated in Figure 1, PLMs can directly generate reasonable label words (e.g., "positive") for a sentence $<S>$ by answering the query of the "[MASK]" token, when fed a prompt-wrapped sequence (e.g., "<S> is this review positive ? [MASK]"). Thus, we argue that the representations of the "[MASK]" token derived by task-specific templates can be viewed as virtual semantic prototypes, which can be obtained without using label information (Lan et al., 2021) and clustering algorithms (Li et al., 2020b; Caron et al., 2020). Besides the commonly-used templates, we manually convert each template to its negation, and use them to induce negative prototypes. Back to Figure 1, with the template "<S> is not a [MASK] one", the word "bad" can be derived from PLMs.

In this paper, we propose ConPVP (**Con**trastive learning with **P**rompt-derived **V**irtual Semantic **P**rototypes) (ConPVP) for unsupervised sentence representation learning. Specifically, given an input sentence, we generate the positive and negative prototypical embeddings by using a task-specific template and its negative counterpart, respectively. We use the contrastive loss to enforce the sentence embedding to be close to its positive prototype, and far apart from the negative prototype as well as the prototypes of other sentences. As illustrated in Figure 2, the issue of local smoothness can be alleviated by exploiting the semantic regularization induced by task-specific prompts, and the sentences with similar semantics are closer. We empirically evaluate our proposed ConPVP on a range of semantic textual similarity tasks, and the experimental results show the substantial improvements compared with strong baselines. Further,the extensive analysis and applications to transfer and clustering tasks confirm the effectiveness and robustness of our ConPVP.

2 Related Work

2.1 Prototypical Contrastive Learning

Recently, prototypical contrastive learning has shown its power in computer vision (Li et al., 2020b; Caron et al., 2020; Sharma et al., 2020) and NLP tasks (Wei et al., 2022; Ding et al., 2021), which discover the underlying semantic structure by clustering the learned embeddings. Compared with them, we propose a more efficient and dedicated method to find prototypes for sentence embeddings, without using clustering algorithms or label information. To the best of our knowledge, we are the first to explore the prototypical contrastive learning in unsupervised sentence representation learning.

2.2 Prompt-based Learning

Prompt-based Learning has become a new paradigm in NLP, bridging the gap between pre-training tasks and downstream tasks (Brown et al., 2020; Schick and Schütze, 2021a; Sanh et al., 2021). It reformulates various NLP tasks as cloze-style questions, and by doing so, the knowledge stored in PLMs can be fully exploited, making PLMs achieve impressive performance in few-shot and zero-shot settings. Along this research line, various types of prompts are explored including discrete and continuous prompts (Gao et al., 2021a; Shin et al., 2020; Hu et al., 2021; Liu et al., 2021; Cui et al., 2021; Si et al., 2021; Li and Liang, 2021; Schick and Schütze, 2021b). In this work, we exploit prompts of different downstream tasks to assign various virtual semantic prototypes to each instance.

2.3 Unsupervised Sentence Embedding

Unsupervised learning has been used to improve the sentence embedding learning (Reimers and

Gurevych, 2019b; Li et al., 2020a; Su et al., 2021; Zhang et al., 2020), and contrastive learning has attracted extensive attention due to the promising performance (Gao et al., 2021b; Giorgi et al., 2021; Wu et al., 2020; Yan et al., 2021; Meng et al., 2021; Carlsson et al., 2021). Wu et al. (2021) augment positive pairs with word repetition and introduce a momentum encoder for negative pairs. Wang et al. (2022) use soft negative samples which have highly similar textual but opposite meaning to the input sentence. Jiang et al. (2022) use a discrete template to obtain sentence embeddings. Unlike these studies, we introduce prototypical contrastive learning and implicitly encode semantic structure induced by task-specific prompts into the embedding space, enhancing PLMs' ability of modeling semantic similarity. Furthermore, our prototypical contrastive loss is orthogonal to the instance-wise one, and the performance can be further improved by combining ConPVP with the above studies.

3 Method

In this section, we elaborate the proposed ConPVP, a novel contrastive learning approach implicitly encoding semantic structure into the embedding space. As illustrated in Figure 3, ConPVP is based on the popular SimCSE framework (Gao et al., 2021b) and further leverages the concept of semantic prototypes.

3.1 Prompt-derived Virtual Semantic Prototypes

Semantic prototype is defined as a representative embedding for a group of semantically similar instances (Li et al., 2020b). Given the fact that PLMs are able to perform well on various NLP tasks when provided with suitable task-specific templates (Brown et al., 2020; Sanh et al., 2021; Wei et al., 2021), we can induce the semantic prototypes of sentences from PLMs with the help of prompts. In this work, we construct a template set $\mathcal{T}^+$ using four NLP tasks (i.e., classification, summarization, natural language inference, and sentence embedding), and assign each task 2 templates. Please note that we select the templates without deliberation, and we leave the other choices as future work. Furthermore, we construct another template set $\mathcal{T}^-$, in which the templates are the negative form of those in $\mathcal{T}^+$ and possibly induces semantically opposite response from PLMs. All the templates are illustrated in Table 1.

Basic Templates
Given "<S>" , we assume that "[MASK]"
" <S> " , is this review positive ? [MASK] .
" <S> " , is [MASK] news
" <S> " , is a [MASK] one
" <S> " . In summary : " [MASK] "
By " <S> " they mean [MASK] .
Article " <S> " belongs to a [MASK] topic
This sentence : " <S> " means [MASK] .
Semantically Opposite Templates
" <S> " , is this review negative ? [MASK] .
Without " <S> " , they mean [MASK] .
" <S> " is inconsistent with " [MASK] "
" <S> " is totally different from : " [MASK] "
" <S> " which does not denote [MASK]
" <S> " is not a [MASK] one
This sentence : " <S> " does not mean [MASK] .
Article " <S> " is definitely not about the [MASK] topic

Table 1: Templates for inducing semantic prototypes.

After obtaining the template sets, we convert an input sentence x_i to $\hat{x}_i$=[x_i; $\mathcal{T}_i^+$], where $\mathcal{T}_i^+$ is a template sampled from $\mathcal{T}^+$. We feed $\hat{x}_i$ to a PLM and take the hidden state of the "[MASK]" token $h_{\text{[MASK]}}$ as the positive prototypical embedding c_i^+. In this same way, we generate the negative prototypical embedding c_i^- using a sampled template $\mathcal{T}_i^- \in \mathcal{T}^-$. Notably, unlike the conventional prototypes (Li et al., 2020b; Caron et al., 2020; Sharma et al., 2020), our method of obtaining prototypes may not seem intuitive, since there is no explicit partitioning of the embedding space. In order to distinguish our method from the previous studies, we name the prototypes in this work virtual prototypes.

3.2 Prototypical Contrastive Learning

To obtain the embedding of an anchor sentence x_i, we feed $[x_i; t_1; ...; t_l$;[MASK]] to a PLM to obtain its contextualized representations, where $t_1; ...; t_l$ is a continuous prompt. The representation vector of the "[MASK]" token is taken as the sentence embedding v_i. Given the embedding of the anchor sentence and the corresponding positive and negative prototype embeddings, we integrate them into

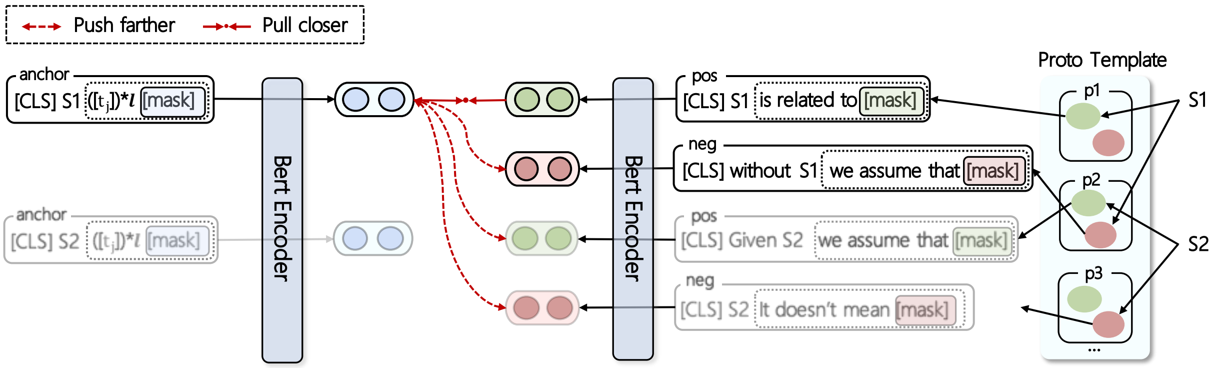

Figure 3: The overall framework of our proposed ConPVP.

PLM	Batch Size	Learning Rate
BERT-base	128	3e-5
BERT-large	128	1e-5
RoBERTa-base	128	1e-5
RoBERTa-large	256	1e-5

Table 2: **Training Settings for STS**.

the InfoNCE based contrastive loss [1]:

$$l_i = -\log \frac{e^{\mathrm{sim}(v_i, c_i^+)/\tau}}{\sum_{k=1}^{N} \left(e^{\mathrm{sim}(v_i, c_k^+)/\tau} + e^{\mathrm{sim}(v_i, c_k^-)/\tau}\right)} \tag{1}$$

where N is the number of sentences in a mini-batch. With this loss function, we pull the embedding of the anchor sentence v_i close to its positive prototypical embedding c_i^+, and push v_i and the irrelevant prototypical embeddings apart.

4 Experiments

To verify the effectiveness of our proposed method, we conduct experiments and empirical analysis on Semantic Textual Similarity (STS) tasks under the unsupervised setting.

4.1 Settings

Following Gao et al. (2021b), we conduct experiments on 7 semantic textual similarity (STS) tasks, including STS 2012-2016 (Agirre et al., 2012, 2013, 2014, 2015, 2016), STS Benchmark (Cer et al., 2017), and SICK-R (Marelli et al., 2014). The similarity scores of sentence pairs in these datasets are labeled from 0 to 5. Our implementation is based on SimCSE [2] (Gao et al., 2021b), and we take BERT-base (Devlin et al., 2019), BERT-large (Devlin et al., 2019), RoBERTa-base (Liu et al., 2019), and RoBERTa-large (Liu et al., 2019) as our backbones[3]. All our experiments are conducted on a NVIDIA V100 GPU. We set the length of the continuous prompt as 4. Following previous studies (Gao et al., 2021b; Wu et al., 2021), we use 1 million sentences randomly sampled from English Wikipedia as training sentences. We train 1 epoch, and evaluate every 125 steps and choose model parameters with highest performance on STS-B development set. The batch size and learning rate are listed in Table 2.

4.2 Main Results

We compare our ConPVP to the recent related methods which are based on instance-wise contrastive learning, including 1) *ConSERT* (Yan et al., 2021) which exploits four data augmentation strategies to construct positive samples; 2) *SimCSE* (Gao et al., 2021b) which directly uses Dropout to generate positive pairs; 3) *ESimCSE* (Wu et al., 2021) which introduces word repetition augmented positive pairs and momentum negative pairs; 4) *PromptBERT* (Jiang et al., 2022) which reformulates the sentence embeddings task as a prompt-based learning paradigm.

For fair comparison, we report the best performance from 4 runs in Table 3. Compared with *SimCSE*, *ConPVP* brings significant improvements across the board. Specifically, *ConPVP* achieves average improvements of 2.60, 1.60, 2.61, and 1.28 points over BERT-base, BERT-large, RoBERTa-base and RoBERTa-large, respectively, showing the superiority of our prototypical contrastive method.

[1] In practice, in order to reduce the influence of the template, we follow (Jiang et al., 2022) to use debiased sentence embeddings during training. Please ref (Jiang et al., 2022) for more detailed.

[2] https://github.com/princeton-nlp/SimCSE

[3] https://github.com/huggingface/transformers

Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
				BERT-base				
ConSERT †	64.64	78.49	69.07	79.72	75.95	73.97	67.31	72.74
SimCSE †	68.40	82.41	74.38	80.91	78.56	76.85	72.23	76.25
ESimCSE †	**73.40**	83.27	77.25	82.66	78.81	80.17	72.30	78.27
PromptBERT †	71.56	84.58	76.98	**84.47**	**80.60**	**81.60**	69.87	78.54
ConPVP	71.72	**84.95**	**77.68**	83.64	79.76	80.82	**73.38**	**78.85**
				BERT-large				
ConSERT †	70.69	82.96	74.13	82.78	76.66	77.53	70.37	76.45
SimCSE †	70.88	84.16	76.43	84.50	79.76	79.26	73.88	78.41
ESimCSE †	**73.21**	85.37	77.73	84.30	78.92	80.73	**74.89**	79.31
PromptBERT	71.55	**86.83**	**78.63**	85.10	79.79	**82.20**	72.19	79.47
ConPVP	72.63	86.68	78.14	**85.50**	**80.13**	82.18	74.79	**80.01**
				RoBERTa-base				
SimCSE †	70.16	81.77	73.24	81.36	80.65	80.22	68.56	76.57
ESimCSE †	69.90	82.50	74.68	83.19	80.30	80.99	70.54	77.44
PromptBERT †	**73.94**	**84.74**	**77.28**	**84.99**	**81.74**	81.88	69.50	79.15
ConPVP	73.20	83.22	76.24	83.37	81.49	**82.18**	**74.59**	**79.18**
				RoBERTa-large				
SimCSE †	72.86	83.99	75.62	84.77	81.80	81.98	71.26	78.90
EsimCSE †	73.20	84.93	76.88	84.86	81.21	82.79	72.27	79.45
PromptBERT	72.89	**86.44**	**78.10**	**85.09**	79.37	81.52	70.85	79.18
ConPVP	**74.75**	84.09	77.88	83.13	**83.44**	**83.64**	**74.31**	**80.18**

Table 3: **Experimental results on unsupervised STS tasks**. Methods with † denote that we directly report the scores from corresponding paper, and others are from our implementation. We run 4 times with different random seeds and report the best **Avg.** for fair comparison.

Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
				RoBERTa-large				
SimCSE	$69.36_{\pm 1.16}$	$82.39_{\pm 0.70}$	$74.33_{\pm 1.06}$	$83.03_{\pm 1.34}$	$81.19_{\pm 0.45}$	$81.10_{\pm 0.77}$	$70.17_{\pm 0.91}$	$77.37_{\pm 0.88}$
PromptBERT	$72.00_{\pm 1.16}$	$83.54_{\pm 1.85}$	$77.05_{\pm 1.02}$	$83.32_{\pm 1.15}$	$80.82_{\pm 0.98}$	$82.54_{\pm 0.79}$	$70.31_{\pm 0.82}$	$78.51_{\pm 0.77}$
ConPVP	$\mathbf{74.57}_{\pm 0.54}$	$\mathbf{83.62}_{\pm 0.59}$	$\mathbf{77.77}_{\pm 0.29}$	$83.18_{\pm 0.85}$	$\mathbf{82.85}_{\pm 0.37}$	$\mathbf{82.85}_{\pm 0.46}$	$\mathbf{74.47}_{\pm 0.44}$	$\mathbf{79.90}_{\pm 0.32}$
w/ manual	$72.61_{\pm 1.26}$	$82.31_{\pm 0.72}$	$77.19_{\pm 0.55}$	$\mathbf{84.85}_{\pm 1.00}$	$81.48_{\pm 0.44}$	$82.31_{\pm 0.48}$	$73.50_{\pm 0.54}$	$79.18_{\pm 0.49}$
w/o c^-	$73.80_{\pm 0.93}$	$82.38_{\pm 0.58}$	$76.72_{\pm 0.50}$	$82.53_{\pm 0.77}$	$81.94_{\pm 0.54}$	$82.32_{\pm 0.31}$	$69.75_{\pm 0.58}$	$78.49_{\pm 0.33}$
w/o c^+ & c^-	$72.81_{\pm 0.75}$	$81.51_{\pm 0.63}$	$74.94_{\pm 0.77}$	$79.83_{\pm 0.81}$	$80.50_{\pm 0.55}$	$81.06_{\pm 0.40}$	$70.30_{\pm 0.43}$	$77.28_{\pm 0.25}$

Table 4: **Ablation Study**. We run each experiment 4 times with different random seeds and report mean and standard deviation.

Besides, *ConPVP* surpasses *ConSERT* and *ESimCSE*, which carefully design positive samples with various textual data augmentation. This demonstrates that although the textual data augmentation can provide different views of the anchor, these methods based on it still suffers the local smooth problem. In contrast, our model shows that textual data augmentation is possibly unnecessary, and the improvement can be achieved by encoding more structural information into the embedding space, e.g., finding semantic prototypes. Finally, our *ConPVP* achieves consistently better performance than *PromptBERT*, demonstrating the effectiveness of the proposed prototypical contrastive loss.

4.3 Ablation Study

To analyze the impact of different components of ConPVP, we investigate the following three variants: 1) *ConPVP w/ manual*, where we obtain the anchor sentence embeddings with the searched discrete templates from Jiang et al. (2022); 2) *ConPVP w/o c^-*, where we remove the negative prototypes

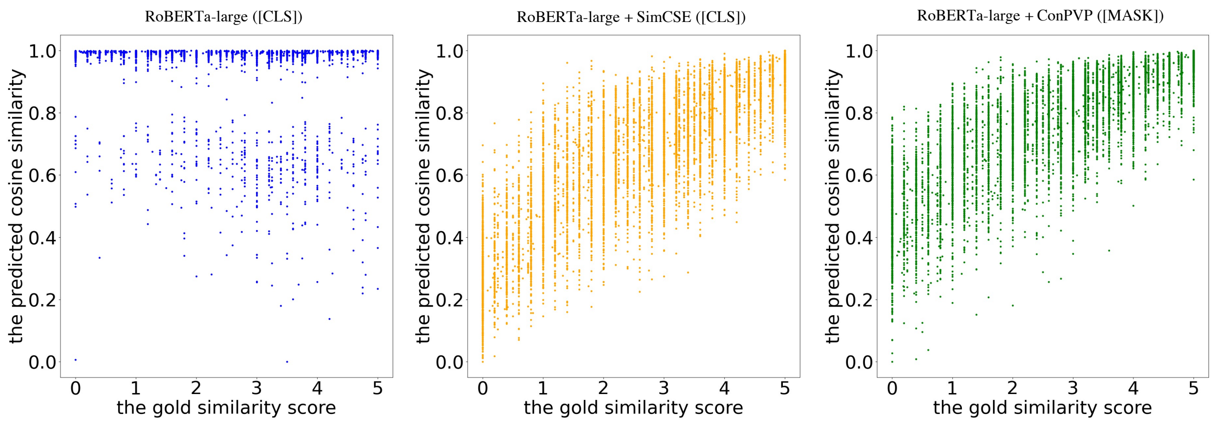

Figure 4: **Distribution of predicted cosine similarity.** The correlation diagram between the gold similarity scores (x-axis) and model predicted cosine similarity scores (y-axis) on the STS-B dataset. We scale the predicted scores to 0 to 1.

in the prototypical contrastive loss; 3) *ConPVP w/o* c^+ *&* c^-, which is equivalent to SimCSE but uses continuous prompts for sentence embeddings. Notably, *ConPVP w/o* c^+ *&* c^- is also a variant of PromptBERT (Jiang et al., 2022), where the discrete templates are replaced by the continuous ones. We take RoBERTa-large as the backbone.

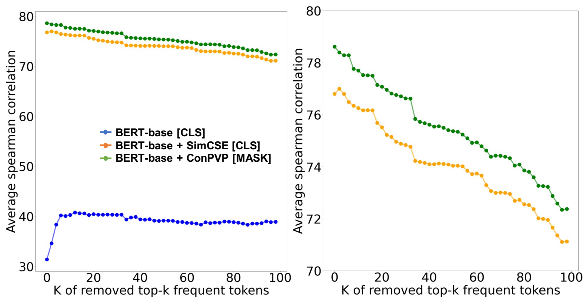

Figure 5: **Analysis of embedding space.** The average spearman correlation on STS tasks w.r.t the number of removed top-k frequent tokens. The frequency of each token is calculated through the test split of the STS Benchmark dataset.

The results on STS tasks are listed in Table 4 and the conclusions are as follows: 1) *ConPVP* obtains better results against *ConPVP w/ manual*. This may be because one manually-designed prompt cannot fit different PLMs and training strategies at the same time, and continuous prompts are more flexible and effective in comparison. Besides, the improvement of *ConPVP w/ manual* over *PromptBERT* validates the advantage of the prototypical contrastive loss. 2) Removing the negative prototypes (i.e., *ConPVP w/o* c^-) leads to a performance degradation of 1.41 point against *ConPVP*. The underlying reason is that the negative prototypes here serve as a type of hard negatives—the semantics of the negative prototypes are essentially different from the positive prototypes but the prompts used to induce them are similar in text. 3) We find that *ConPVP w/o* c^+ *&* c^- does not give an improvement against *SimCSE*. These observations show that the gain of our method entirely comes from the cooperation between the prompt-derived virtual prototypes and the prototypical contrastive loss, rather than the usage of the prompt-based sentence embeddings.

4.4 Distribution of Cosine Similarity

In this section, we investigate the similarity distributions learned by different methods. As shown in Figure 4, the native sentence representations of *RoBERTa-large* suffer from the collapse issue (Chen and He, 2021), and therefore we get high similarity scores for all sentence pairs. By contrast, both *ConPVP* and *SimCSE* alleviate the collapse issue, and the predicted cosine similarity scores for positive pairs of *ConPVP* are more certain. For example, for the positive pairs whose similarity scores range from 4 to 5, the scores predicted by *ConPVP* (0.6 to 1.0) is more concentrated than the scores predicted by *SimCSE* (0.4 to 1.0).

4.5 Analysis of Embedding Space

Previous studies indicated that the collapse issue is mainly due to anisotropy of the learned embedding space, which is sensitive to token frequency (Yan et al., 2021; Jiang et al., 2022). We follow Yan et al. (2021) to remove the embeddings of K most frequent tokens and explore the relation between the number of removed tokens and the average

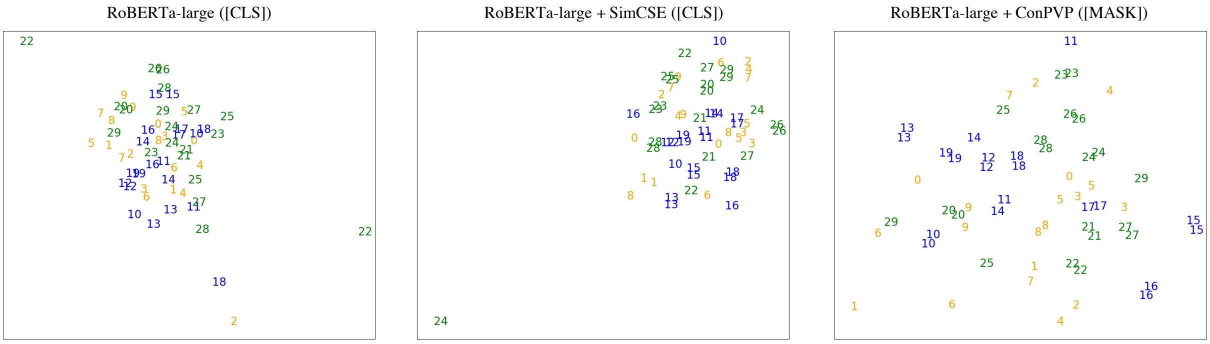

Figure 6: **Visualization of learned embeddings.** We visualize 10 sentence pairs whose similarity scores are 0 in orange (ids from 0 to 9), 10 pairs whose similarity scores are 3 in blue (ids from 10 to 19), and 10 pairs whose similarity scores are 5 in green (ids from 20 to 29). The sentences are sampled from the STS-B test set.

Model	**MR**	**CR**	**SUBJ**	**MPQA**	**SST**	**TREC**	**MRPC**	**Avg.**
GloVe †	77.25	78.30	91.17	87.85	80.18	83.00	72.87	81.52
Skip-thought ♡	76.50	80.10	93.60	87.10	82.00	92.20	73.00	83.50
IS-BERT ♡	81.09	87.18	94.96	88.75	85.96	88.64	74.24	85.83
				RoBERTa-base				
SimCSE †	81.04	87.74	93.28	86.94	86.60	84.60	73.68	84.84
SimCSE	81.75$_{\pm 0.19}$	87.23$_{\pm 0.08}$	93.18$_{\pm 0.13}$	87.13$_{\pm 0.06}$	86.98$_{\pm 0.39}$	85.40$_{\pm 0.71}$	73.78$_{\pm 0.11}$	85.06$_{\pm 0.13}$
Ours	**82.44**$_{\pm 0.17}$	**88.30**$_{\pm 0.16}$	**93.20**$_{\pm 0.11}$	**88.74**$_{\pm 0.06}$	**87.70**$_{\pm 0.07}$	**87.33**$_{\pm 0.25}$	**76.15**$_{\pm 0.19}$	**86.27**$_{\pm 0.11}$
				RoBERTa-large				
SimCSE †	82.74	87.87	93.66	88.22	88.58	92.00	69.68	86.11
SimCSE	83.17$_{\pm 0.41}$	88.46$_{\pm 0.43}$	93.73$_{\pm 0.10}$	88.33$_{\pm 0.10}$	88.52$_{\pm 0.29}$	91.40$_{\pm 0.71}$	71.34$_{\pm 1.17}$	86.42$_{\pm 0.23}$
ConPVP	**85.65**$_{\pm 0.28}$	**90.73**$_{\pm 0.32}$	**94.13**$_{\pm 0.13}$	**90.03**$_{\pm 0.23}$	**89.81**$_{\pm 0.29}$	**93.40**$_{\pm 0.16}$	**76.47**$_{\pm 0.29}$	**88.60**$_{\pm 0.14}$

Table 5: **Experimental results on Transfer tasks with** *RoBERTa-base* **and** *RoBERTa-large* **backbones**. †: results from Gao et al. (2021b). ♡: results from Zhang et al. (2020). We run 4 times with different random seeds and report the average accuracy and standard deviation.

spearman correlation on STS tasks.

From Figure 5, we can observe that the performance of native *BERT-base* and *SimCSE* improves when removing the most frequent tokens. By contrast, *ConPVP* achieves its best performance without removing any tokens, showing that our approach reshapes the BERT's original embedding space, reducing the influence of common tokens on sentence representations. In addition, the performance of both *SimCSE* and *ConPVP* drops as the number of removed tokens increases but *ConPVP* performs significantly better, demonstrating the robustness of *ConPVP* to incomplete input.

4.6 Visualization of Learned Embeddings

We visualize a few variants of *RoBERTa-large* sentence embeddings to grasp an intuition on the effectiveness of our method. Specifically, we sample 3 groups of samples from the STS-B test set, and the similarity score of each group is 0 (orange), 3 (blue), and 5 (green), respectively. Each group has 10 sentence pairs. We visualize their embeddings generated by different models using t-SNE (van der Maaten and Hinton, 2008) in Figure 6.

Due to the collapse issue, the sentence embeddings obtained from *RoBERTa-large [CLS]* cluster together whether they are similar or not. For SimCSE, the sentence embeddings of the positive pairs are well-clustered. However, the sentences pairs with similarity scores of 3 or 5 are very close in the embedding space. In contrast, the embeddings learned by our ConPVP are more discriminative, forming more separated clusters (e.g., the sentence pairs in green are more clustered than those in blue, while the pairs in orange are more dispersed).

5 Application to Transfer Learning Tasks

We evaluate the quality of the sentence embeddings learned by ConPVP on transfer learning tasks, in-

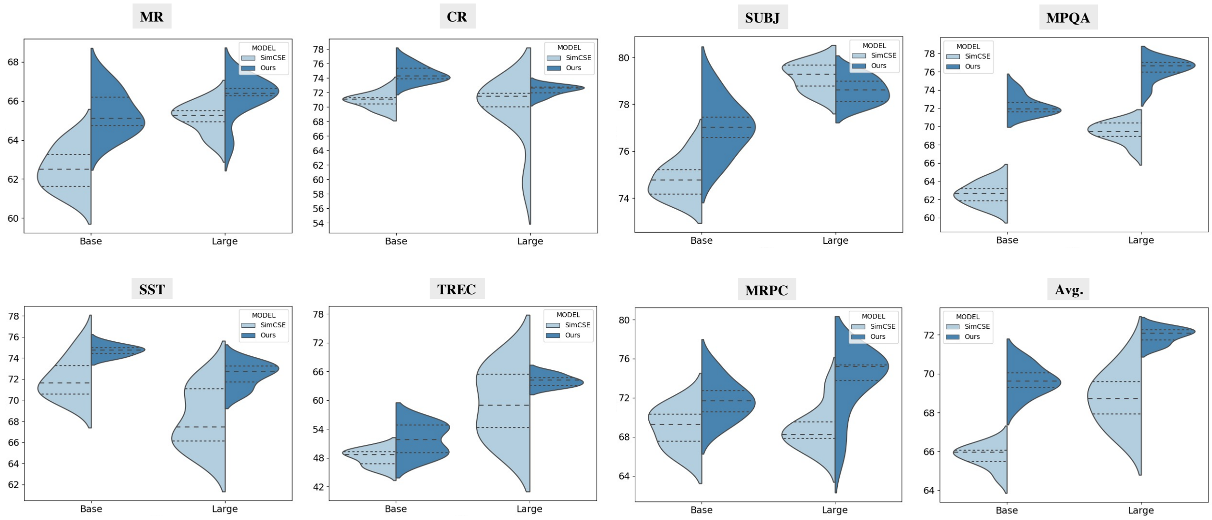

Figure 7: **Few-shot learning evaluation on Transfer tasks with *RoBERTa-base* and *RoBERTa-large* as the backbones**. For each task, we randomly sample 16 labeled instances per class and draw violin plots of the performance of 10 runs with different random seeds.

Model	AG	Bio	Go-S	G-T	G-TS	SS	SO	Tweet	Avg.
					BERT-base				
BERT	79.56	32.46	54.35	47.12	61.61	64.04	21.87	45.35	50.80
SimCSE ♣	74.36	35.89	58.90	57.28	65.03	64.32	50.57	54.28	57.58
ConPVP	**77.21**	**41.84**	**61.98**	**59.87**	**67.56**	**73.28**	**73.06**	**56.06**	**63.86**
					BERT-large				
BERT	**83.13**	30.52	56.34	46.11	61.51	66.54	26.10	44.20	51.81
SimCSE ♣	80.23	**43.47**	61.87	**61.05**	65.78	68.97	**68.03**	55.08	63.06
ConPVP	82.50	41.26	**63.82**	58.87	**68.34**	**74.39**	66.59	**57.34**	**64.14**

Table 6: **Clustering accuracy reported on short text clustering datasets with *BERT-base* and *BERT-large* as the backbones**. ♣: results evaluated on the checkpoints provided by (Gao et al., 2021b). We report the clustering accuracy averaged over 10 independent runs.

cluding MR (Pang and Lee, 2005), CR (Hu and Liu, 2004), SUBJ (Pang and Lee, 2004), MPQA (Wiebe et al., 2005), SST-2 (Socher et al., 2013), TREC (Li and Roth, 2002) and MRPC (Dolan et al., 2004). A logistic regression classifier is trained using frozen sentence embeddings produced by different methods. We follow default configurations from SentEval (Conneau and Kiela, 2018). In addition, based on the principle that good representations can be transferred well with limited supervision and fine-tuning, we extend the evaluation to few-shot setting and follow (Zhang et al., 2021) to uniformly sample 16 labeled instances per class for each task.

Table 5 presents the results under the full data setting. As we can see, the performance gap between *ConPVP* and *SimCSE* is significant and consistent. Furthermore, we can observe more obvious gap under the few-shot setting (Figure 7). The results reveal the robustness and effectiveness of our approach under the data scarcity scenarios, which is important in real-world applications.

6 Application to Clustering Tasks

We follow Zhang et al. (2021) to consider 8 benchmark datasets for short text clustering, including SearchSnippets (SS) (Phan et al., 2008), StackOverflow (SO) (Xu et al., 2017), Biomedical (Bio) (Xu et al., 2017), AgNews (AG) (Zhang and LeCun, 2015), Tweet (Yin and Wang, 2016) and GoogleNews (G-T, G-S, G-TS) (Yin and Wang, 2016). We follow default settings of (Zhang et al., 2021) and use BERT-base and BERT-large as the backbones. We run K-Means (Pedregosa et al., 2011) on the sentence embeddings and report the clustering accuracy averaged over 10 independent runs. As illustrated in Table 6, in comparison with *SimCSE*, *ConPVP* obtains an averaged improvement of 1.21 and 2.18, respectively, which validates our motivation in leveraging the implicit grouping effect of the prompt-derived semantic prototypes to encode more semantic structure into representations.

7 Conclusion

In this work, we take the first step to explore the prototypical contrastive learning on unsupervised sentence embedding learning, and consider more semantic views for each instance than the recent instance-wise contrastive methods. In particular, we make use of the prompting in PLMs to generate the positive and negative prototypical embeddings with task-specific templates. The experiments and extensive analysis validate the effectiveness and robustness of our ConPVP.

8 Limitations

We only tried 16 task specific prompts in this paper, which is possibly sub-optimal to induce semantic prototypes. Besides, the usage of prompts reduces the maximum effective lengths that the pretrained language models can process.

Acknowledgements

We would like to thank anonymous reviewers for their suggestions and comments.

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.

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. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Fredrik Carlsson, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Hellqvist, and Magnus Sahlgren. 2021. Semantic re-tuning with contrastive tension. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2020. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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.

Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 15750–15758. Computer Vision Foundation / IEEE.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

Leyang Cui, Yu Wu, Jian Liu, Sen Yang, and Yue Zhang. 2021. Template-based named entity recognition using BART. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1835–1845, Online. 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.

Ning Ding, Xiaobin Wang, Yao Fu, Guangwei Xu, Rui Wang, Pengjun Xie, Ying Shen, Fei Huang, Hai-Tao Zheng, and Rui Zhang. 2021. Prototypical representation learning for relation extraction. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *COLING 2004, 20th International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2004, Geneva, Switzerland*.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021a. Making pre-trained language models better few-shot learners. 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 3816–3830, Online. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021b. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. DeCLUTR: Deep contrastive learning for unsupervised textual representations. 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 879–895, Online. Association for Computational Linguistics.

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.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004*, pages 168–177. ACM.

Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Juanzi Li, and Maosong Sun. 2021. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. *arXiv preprint arXiv:2108.02035*.

Ting Jiang, Shaohan Huang, Zihan Zhang, Deqing Wang, Fuzhen Zhuang, Furu Wei, Haizhen Huang, Liangjie Zhang, and Qi Zhang. 2022. Promptbert: Improving bert sentence embeddings with prompts. *arXiv preprint arXiv:2201.04337*.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3294–3302.

Jinhe Lan, Qingyuan Zhan, Chenhao Jiang, Kunping Yuan, and Desheng Wang. 2021. CLLD: contrastive learning with label distance for text classificatioin. *CoRR*, abs/2110.13656.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020a. 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.

Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi. 2020b. Prototypical contrastive learning of unsupervised representations. *Proceedings of ICLR 2021*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. 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 4582–4597, Online. Association for Computational Linguistics.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *19th International Conference on Computational Linguistics, COLING 2002, Howard International House and Academia Sinica, Taipei, Taiwan, August 24 - September 1, 2002*.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021. Gpt understands, too. *arXiv preprint arXiv:2103.10385*.

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. *arXiv preprint arXiv:1907.11692*.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

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).

Yu Meng, Chenyan Xiong, Payal Bajaj, Paul Bennett, Jiawei Han, Xia Song, et al. 2021. Coco-lm: Correcting and contrasting text sequences for language model pretraining. *Advances in Neural Information Processing Systems*, 34.

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.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 21-26 July, 2004, Barcelona, Spain*, pages 271–278. ACL.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL 2005, 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, USA*, pages 115–124. The Association for Computer Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12:2825–2830.

Xuan Hieu Phan, Minh Le Nguyen, and Susumu Horiguchi. 2008. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th International Conference on World Wide Web, WWW 2008, Beijing, China, April 21-25, 2008*, pages 91–100. ACM.

Nils Reimers and Iryna Gurevych. 2019a. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019b. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M. Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. 2021. Multitask prompted training enables zero-shot task generalization. *CoRR*, abs/2110.08207.

Timo Schick and Hinrich Schütze. 2021a. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021b. Few-shot text generation with natural language instructions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 390–402, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vivek Sharma, Makarand Tapaswi, M. Saquib Sarfraz, and Rainer Stiefelhagen. 2020. Clustering based contrastive learning for improving face representations. In *15th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2020, Buenos Aires, Argentina, November 16-20, 2020*, pages 109–116. IEEE.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Jinghui Si, Xutan Peng, Chen Li, Haotian Xu, and Jianxin Li. 2021. Generating disentangled arguments with prompts: A simple event extraction framework that works. *arXiv preprint arXiv:2110.04525*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1631–1642. ACL.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Dong Wang, Ning Ding, Piji Li, and Haitao Zheng. 2021a. CLINE: Contrastive learning with semantic negative examples for natural language understanding. 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 2332–2342, Online. Association for Computational Linguistics.

Hao Wang, Yangguang Li, Zhen Huang, Yong Dou, Lingpeng Kong, and Jing Shao. 2022. Sncse: Contrastive learning for unsupervised sentence embedding with soft negative samples. *arXiv preprint arXiv:2201.05979*.

Kexin Wang, Nils Reimers, and Iryna Gurevych. 2021b. TSDAE: using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 671–688. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned language models are zero-shot learners. *CoRR*, abs/2109.01652.

Yinyi Wei, Tong Mo, Yongtao Jiang, Weiping Li, and Wen Zhao. 2022. Eliciting knowledge from pretrained language models for prototypical prompt verbalizer. *CoRR*, abs/2201.05411.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Lang. Resour. Evaluation*, 39(2-3):165–210.

Xing Wu, Chaochen Gao, Liangjun Zang, Jizhong Han, Zhongyuan Wang, and Songlin Hu. 2021. Esimcse: Enhanced sample building method for contrastive learning of unsupervised sentence embedding. *CoRR*, abs/2109.04380.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. Clear: Contrastive learning for sentence representation. *arXiv preprint arXiv:2012.15466*.

Jiaming Xu, Bo Xu, Peng Wang, Suncong Zheng, Guanhua Tian, and Jun Zhao. 2017. Self-taught convolutional neural networks for short text clustering. *Neural Networks*, 88:22–31.

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.

Jianhua Yin and Jianyong Wang. 2016. A model-based approach for text clustering with outlier detection. In *32nd IEEE International Conference on Data Engineering, ICDE 2016, Helsinki, Finland, May 16-20, 2016*, pages 625–636. IEEE Computer Society.

Dejiao Zhang, Shang-Wen Li, Wei Xiao, Henghui Zhu, Ramesh Nallapati, Andrew O. Arnold, and Bing Xiang. 2021. Pairwise supervised contrastive learning of sentence representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5786–5798. Association for Computational Linguistics.

Xiang Zhang and Yann LeCun. 2015. Text understanding from scratch. *CoRR*, abs/1502.01710.

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.