Improved Universal Sentence Embeddings with Prompt-based Contrastive Learning and Energy-based Learning

Yuxin Jiang[1,2] **Linhan Zhang**[3] **Wei Wang**[1,2]
[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]The Hong Kong University of Science and Technology
[3]School of Computer Science and Engineering, The University of New South Wales
yjiangcm@connect.ust.hk, linhan.zhang@student.unsw.edu.au, weiwcs@ust.hk

Abstract

Contrastive learning has been demonstrated to be effective in enhancing pre-trained language models (PLMs) to derive superior universal sentence embeddings. However, existing contrastive methods still have two limitations. Firstly, previous works may acquire poor performance under domain shift settings, thus hindering the application of sentence representations in practice. We attribute this low performance to the over-parameterization of PLMs with millions of parameters. To alleviate it, we propose PromCSE (Prompt-based Contrastive Learning for Sentence Embeddings), which only trains small-scale *Soft Prompt* (i.e., a set of trainable vectors) while keeping PLMs fixed. Secondly, the commonly used NT-Xent loss function of contrastive learning does not fully exploit hard negatives in supervised learning settings. To this end, we propose to integrate an Energy-based Hinge loss to enhance the pairwise discriminative power, inspired by the connection between the NT-Xent loss and the Energy-based Learning paradigm. Empirical results on seven standard semantic textual similarity (STS) tasks and a domain-shifted STS task both show the effectiveness of our method compared with the current state-of-the-art sentence embedding models.[1]

1 Introduction

Learning universal sentence embeddings (Kiros et al., 2015; Hill et al., 2016; Conneau et al., 2017; Cer et al., 2018; Reimers and Gurevych, 2019) which convey high-level semantic information without task-specific fine-tuning is a vital research problem in Natural Language Processing (NLP) communities. It could benefit a wide range of applications such as information retrieval, question answering, etc (Logeswaran and Lee, 2018). Recently, fine-tuning Pre-trained Language Models (PLMs) (Devlin et al., 2019) with *contrastive learning*, which aims to pull semantically close samples together and push apart dissimilar samples, has achieved extraordinary success in learning universal sentence representations (Liu et al., 2021a; Giorgi et al., 2021b; Kim et al., 2021; Gao et al., 2021; Chuang et al., 2022). In these works, positive pairs are formed via data augmentation or supervised datasets, whereas negative pairs are derived from different sentences within the same mini-batch. Then contrastive learning objective like normalized temperature-scaled cross-entropy loss (NT-Xent) (Chen et al., 2020a; Gao et al., 2021) is used for optimizing the model parameters. As a typical example, SimCSE (Gao et al., 2021) uses the standard dropout as augmentation for constructing positive pairs and achieves extraordinarily strong performance on seven standard Semantic Textual Similarity (STS) tasks.

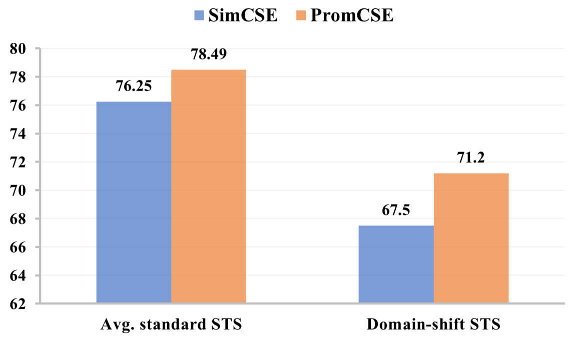


Figure 1: The performance comparison between unsupervised SimCSE and unsupervised PromCSE. Both models are trained on 1 million unlabeled sentences from English Wikipedia.

Though effective, existing contrastive methods for learning sentence representations still have two limitations. *Firstly*, since universal sentence embeddings are often trained on a large corpus and used off-the-shelf on a diverse range of tasks, such domain shifts are commonplace and may pose challenges to the performance. As Figure 1 shows,

[1]Our code is publicly avaliable at `https://github.com/YJiangcm/PromCSE`

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 3021–3035

SimCSE's performance drops significantly when applied on a domain-shifted STS task (Parekh et al., 2021), where the texts are image captions. Such non-robustness of large PLMs towards domain shifts has also been observed in other studies. Ma et al. (2019); Lester et al. (2021) found that tuning PLMs with millions of parameters may result in overfitting to the training data distribution and hence vulnerability to domain shifts. *Secondly*, the commonly used NT-Xent loss function in supervised sentence embedding models does not fully exploit the hard negatives. Moreover, recent studies (Wang et al., 2018; Deng et al., 2019) have shown that the softmax-based loss is insufficient to acquire the discriminating power. Thus, NT-Xent loss in supervised models may not separate positives and hard negatives sufficiently.

In this paper, we propose two techniques to address the above-mentioned limitations. Firstly, we propose the **Prom**pt-based **C**ontrastive Learning for **S**entence **E**mbeddings (**PromCSE**) to alleviate the domain shift problem, inspired by prompt tuning (Lester et al., 2021; Li and Liang, 2021). Specifically, we modify SimCSE by freezing the entire pre-trained model and add multi-layer learnable *Soft Prompt*, which is simple yet achieves a good balance between the expressiveness and the robustness to distributional changes. Secondly, we show that the contrastive learning framework under NT-Xent loss (Chen et al., 2020b) could be seen as an instance of Energy-Based Learning (Hinton, 2002; LeCun et al., 2006; Ranzato et al., 2007). Inspired by this connection, we propose an Energy-based Hinge (EH) loss to supplement NT-Xent loss under supervised settings, which enhances the pairwise discriminative power by explicitly creating an energy gap between positive pairs and the hardest negative pairs. We performed extensive experiments using the seven commonly used STS tasks and another out-of-domain STS task. For the same-domain setting, the unsupervised PromCSE can outperform SimCSE by around 2.2 points and is on par with the current state-of-the-art (SOTA) sentence embedding method on the seven standard STS tasks. For the out-of-domain setting, the proposed unsupervised PromCSE can achieve 3.7 absolute points improvements over SimCSE and even 1.2 absolute points improvements over the current SOTA method, which demonstrates its robustness to domain shifts. Moreover, we also demonstrate that the EH loss can improve supervised SimCSE and PromCSE consistently over multiple pre-trained backbone models, achieving state-of-the-art results among supervised sentence representation learning methods.

Our contributions are summarized as follows:

- We identified two limitations of the SOTA methods for both unsupervised and supervised universal sentence representation learning in their robustness to domain shifts and the formulation of their loss functions.
- We propose a multi-layer, prompt-based solution, dubbed PromCSE, as a robust framework for learning sentence embeddings in both the supervised and unsupervised settings.
- We proposed the addition of an Energy-based loss function term to the above contrastive learning framework which can further boost the performance of supervised sentence embeddings.
- Empirical results on seven standard STS tasks and one domain-shifted STS task both verify the effectiveness of our proposed method.

2 Related Work

2.1 Sentence Representation Learning

Learning universal sentence representations has been studied extensively in prior works, roughly categorized into supervised (Conneau and Kiela, 2018; Cer et al., 2018) and unsupervised approaches (Hill et al., 2016; Li et al., 2020). Supervised methods train the sentence encoder on datasets with annotations like the supervised Natural Language Inference (NLI) tasks (Cer et al., 2018; Reimers and Gurevych, 2019). Unsupervised approaches consider deriving sentence embeddings without annotated data, *e.g.*, average GloVe embeddings (Pennington et al., 2014), FastSent (Hill et al., 2016) and Quick-Thought (Logeswaran and Lee, 2018). To leverage the rich semantic information implicitly learned by PLMs, recent works have proposed several technics to mitigate the anisotropy issue (Ethayarajh, 2019; Li et al., 2020) of PLMs. Post-processing methods like BERT-flow (Li et al., 2020) and BERT-whitening (Su et al., 2021) attempt to regularize the semantic space of sentences. Contrastive learning approaches learn sentence embeddings by creating semantically close augmentations and pulling these representations to be closer than representations of random negative examples, which have achieved significant performance improvement (Yan et al., 2021; Liu et al., 2021a;

Giorgi et al., 2021a; Gao et al., 2021; Jiang et al., 2022; Shou et al., 2022; Zhou et al., 2022; Zhang et al., 2022; Chuang et al., 2022).

2.2 Language Model Prompting

The language model prompting has emerged with the introduction of GPT-3 (Brown et al., 2020), which demonstrates promising few-shot performance. Previous works design various discrete prompts manually for specific tasks such as knowledge extraction (Petroni et al., 2019). To reduce the tedious process of prompt selection, works like (Schick and Schütze, 2020a,b; Shin et al., 2020) focus on automatically searching discrete prompts. However, the prompt search over discrete space is time-consuming and sub-optimal due to the continuous nature of neural networks. To solve these issues, (Lester et al., 2021; Li and Liang, 2021; Zhong et al., 2021; Liu et al., 2021b) propose to use soft prompts, which are sets of trainable vectors in the frozen PLMs. These vectors allow the optimization of the downstream tasks in an end-to-end manner. As shown in (Lester et al., 2021), PLMs with *Soft Prompts* can often perform better in domain-shift settings.

2.3 Energy-based Learning

Energy-based Learning provides a common theoretical framework for many learning models, both probabilistic and non-probabilistic (Hinton, 2002; LeCun et al., 2006; Ranzato et al., 2007). Energy-Based Models (EBMs) involve four key components: a scalar *energy* function to measure the degree of compatibility between each configuration of the variables; the *inference* algorithm consisting in setting the value of observed variables and finding values of the remaining variables that minimize the energy; the *loss* function which measures the quality of the available energy functions using the training set; the *learning* algorithm consisting in finding an energy function that associates low energies to correct values of the remaining variables, and higher energies to incorrect values. So far, EBMs have been applied in sparse representation learning (Ranzato et al., 2006), language modeling (Mnih and Teh, 2012), text generation (Deng et al., 2020), etc.

3 Methodology

In this section, we first present *PromCSE*, a prompt-based contrastive learning framework for both unsupervised and supervised sentence representation learning in Section 3.1. Then we demonstrate that the contrastive learning framework under NT-Xent loss is an instance of Energy-based Learning in Section 3.2. Eventually, inspired by Energy-based Learning, we design an Energy-based Hinge loss to supplement NT-Xent loss when hard negatives are available in supervised datasets in Section 3.3.

3.1 Prompt-based Contrastive Learning

Our prompt-based contrastive learning framework consists of two steps. Firstly, an encoder is built by prepending *Soft Prompt* at *each* layer of the PLM to acquire the sentence representation. Then we optimize the sentence embedding vector space based on the contrastive learning objective.

Sentence Encoder with Soft Prompt Fine-tuning is the prevalent way to adapt Transformer-based PLMs as encoders to obtain universal sentence representations. However, model tuning may be over-parameterized and more prone to overfit the training data, to the detriment of similar tasks in different domains.

As an alternative paradigm, prompt tuning (Lester et al., 2021; Li and Liang, 2021) that conditions a frozen PLM with *Soft Prompt* (i.e., a sequence of continuous vectors prepended to the input of PLMs) has been demonstrated to be competitive with full model tuning while conferring benefits in robustness to domain shifts. By freezing the core language model parameters, prompt tuning prevents the model from modifying its general understanding of language. Instead, prompt representations indirectly modulate the representation of the input. This reduces the model's ability to overfit a dataset by memorizing specific lexical cues and spurious correlations. Motivated by this, we propose to utilize prompt tuning for universal sentence representations. During training, we only update the parameters of soft prompts and fix all PLMs parameters.

Different from (Lester et al., 2021) which only adds *Soft Prompt* at the input layer, we prepend a sequence of trainable vectors $P^j = \{\mathbf{p}_1^k, ..., \mathbf{p}_l^k\}$ at *each* transformer layer inspired by (Liu et al., 2021b). Then the i^{th} hidden states at the j^{th} layer $\mathbf{h}_i^j$ in the Transformers (Vaswani et al., 2017) are defined as follows:

$$\mathbf{h}_i^j = \begin{cases} \mathbf{e}_i^j, & j = 0 \wedge i > k \\ \mathbf{p}_i^j, & i \leq k \\ Trans(\mathbf{h}^{j-1})_i, & \text{otherwise} \end{cases} \quad (1)$$

where $Trans()$ denotes the forward function of the Transformer block layer and $\mathbf{e}_i$ denotes the fixed token embedding vector at the input layer. Compared with (Lester et al., 2021), this enables gradients to be backward updated at each layer and better complete the learning tasks. During the training, sentences are fed into the frozen PLM with the prepended *Soft Prompt*, and we add an MLP layer over the $[CLS]$ hidden state from the last layer of PLM to obtain the sentence embeddings.

Contrastive Learning Objective We use the most widely adopted training objective NT-Xent loss (Chen et al., 2020a; Gao et al., 2021), which has been applied in previous sentence and image representation learning methods. Given a set of paired sentences $\mathcal{D} = \{(X_i, X_i^+)\}_{i=1}^m$ where X_i and X_i^+ are semantically close, we regard X_i^+ as positive of X_i and other sentences in the same mini-batch as negatives. Let $\mathbf{h}_i$ and $\mathbf{h}_i^+$ denote the sentence embeddings of X_i and X_i^+, then NT-Xent loss for a single sample in a mini-batch of size N can be formulated as follows:

$$\mathcal{L}_{CL} = -\log \frac{e^{sim(\mathbf{h}_i, \mathbf{h}_i^+)/\tau}}{\sum_{j=1}^{N} e^{sim(\mathbf{h}_i, \mathbf{h}_j^+)/\tau}} \tag{2}$$

where τ is a temperature hyperparameter and $sim(\mathbf{h}_1, \mathbf{h}_2)$ is the cosine similarity function.

We follow SimCSE (Gao et al., 2021) that constructs positive pairs by feeding the same sentence to the sentence encoder twice with diverse dropout masks when only unlabeled text data is available.

3.2 Connecting Contrastive Learning with Energy-based Learning

Given a set of training samples $\mathcal{S} = \{(X_i, Y_i), i = 1 \ldots N\}$ where X and Y are two variables, Energy-Based Models (EBMs) use an scalar *energy function* $E(W, Y_i, X_i)$ indexed by parameter W to measure the compatibility between two variables. Note that small energy values correspond to highly compatible configurations of the variables, while large energy values correspond to highly incompatible configurations. The generalized negative log-likelihood loss of EBMs (LeCun et al., 2006), which stems from a probabilistic formulation of the learning problem in terms of the maximum conditional probability principle, is defined as follows:

$$\mathcal{L}_{nll} = E(W, Y_i, X_i) + \mathcal{F}_\beta(W, \mathcal{Y}, X_i) \tag{3}$$

where $\mathcal{Y}$ is the set of all possible values of Y, $\mathcal{F}$ is the *free energy* of the ensemble $\{E(W, y, X_i), y \in \mathcal{Y}\}$:

$$\mathcal{F}_\beta(W, \mathcal{Y}, X_i) = \frac{1}{\beta} \log \left(\int_{y \in \mathcal{Y}} e^{-\beta E(W, y, X_i)} \right) \tag{4}$$

where β is a positive constant akin to an inverse temperature. Consequently,

$$\mathcal{L}_{nll} \propto -\log \frac{e^{-\beta E(W, Y_i, X_i)}}{\int_{y \in \mathcal{Y}} e^{-\beta E(W, y, X_i)}} \tag{5}$$

Considering X_i and Y_i are both sentences under the *implicit constraint* that X_i and Y_i are positive pairs, we can define the energy function E as

$$E(W, Y_i, X_i) = -sim(f(X_i), f(Y_i)) \tag{6}$$

where f is the sentence encoder parameterized by W. According to Equation (6), the loss in Equation (5) can be rewritten as

$$\mathcal{L}_{nll} \propto -\log \frac{e^{sim(f(X_i), f(Y_i))/\frac{1}{\beta}}}{\int_{y \in \mathcal{Y}} e^{sim(f(X_i), f(y))/\frac{1}{\beta}}} \tag{7}$$

Therefore, we can see that the contrastive loss in Equation (2) can be deemed as a special case of the Energy-based negative log-likelihood loss.

3.3 Energy-based Hinge Loss

NLI datasets (Bowman et al., 2015; Williams et al., 2018) that contain entailment, neutral, and contradiction sentence pairs have shown great success in supervised sentence embedding learning (Conneau et al., 2017; Reimers and Gurevych, 2019). Supervised SimCSE incorporate annotated sentence pairs in contrastive learning by leveraging entailment pairs as positives and extending in-batch negatives with contradiction pairs, namely *hard negatives*. The NT-Xent loss for supervised SimCSE is:

$$\mathcal{L}_{CL} = -\log \frac{e^{sim(\mathbf{h}_i, \mathbf{h}_i^+)/\tau}}{\sum_{j=1}^{N} (e^{sim(\mathbf{h}_i, \mathbf{h}_j^+)/\tau} + e^{sim(\mathbf{h}_i, \mathbf{h}_j^-)/\tau})} \tag{8}$$

where $\mathbf{h}_i, \mathbf{h}_i^+, \mathbf{h}_j^-$ correspond to the embeddings of premise, entailment hypotheses and contradiction hypotheses. Compared with in-batch negatives, hard negatives are more syntactically similar to the anchor, thus making them more likely to be misidentified as positives by the model. In supervised and metric learning literature, it is well-known that hard (i.e., true negative) examples can help guide a learning method to correct its mistakes more quickly (Schroff et al., 2015; Song et al., 2016). However, the softmax-based NT-Xent loss is shown to be insufficient to acquire the discriminating power (Wang et al., 2018; Deng et al., 2019), which may not adequately separate hard negatives and positives. Besides, when the temperature $\tau \to 0^+$, NT-Xent loss degenerates to a triplet loss with a margin of 0 (Wang and Liu, 2021). The small $\tau = 0.05$ used in SimCSE avoids this case but may still cause the sentence representations insufficiently discriminative and, as a result, not sufficiently robust to noise due to the small margin.

To alleviate the above-mentioned problem and inspired by the Energy-based Learning (LeCun et al., 2006), we propose to use the Energy-based Hinge (EH) loss to supplement the original contrastive objective. We first give the following definition:

Definition 1 *Suppose Y is a discrete variable. For a training sample (X_i, Y_i), the* ***most offending incorrect answer*** *$\hat{Y}_i$ is the one that has the lowest energy among all answers that are incorrect:*

$$\hat{Y}_i = \underset{Y \in \mathcal{Y} \wedge Y \neq Y_i}{\arg\min} \; E(W, Y, X_i) \tag{9}$$

Accordingly, the Energy-based Hinge (EH) loss is defined as follows:

$$[m + E(W, Y_i, X_i) - E(W, \hat{Y}_i, X_i)]_+ \tag{10}$$

where $m \geq 0$ is the margin, and $[x]_+ \equiv \max(0, x)$. Combining Equation (6) with Equation (10), we can derive the energy-based hinge loss for sentence embeddings:

$$\mathcal{L}_{EH} = [m + sim(\mathbf{h}_i, \hat{\mathbf{h}}_i) - sim(\mathbf{h}_i, \mathbf{h}_i^+)]_+ \tag{11}$$

The EH loss enhances the pairwise discriminative power by maximizing the decision margin m in the semantic space. During the training, we use the nearest sample among in-batch negatives and hard negatives to approximate the *most offending incorrect answer*; this works empirically well as we observed that it is often the corresponding contradiction hypothesis. Eventually, we can enhance the optimization objective for our **supervised** models with the combination of Equation (8) and (11)

$$\mathcal{L} = \mathcal{L}_{CL} + \lambda \cdot \mathcal{L}_{EH} \tag{12}$$

where λ is a weighting coefficient. We set λ to 10 empirically because the scale of $\mathcal{L}_{EH}$ is around ten smaller than $\mathcal{L}_{CL}$ during training.

4 Experiments

Our experiments are composed of two parts. We first verify the effectiveness of our proposed approach on seven standard STS tasks in Section 4.1. Then we evaluate the domain shift robustness of our approach by testing on a domain shift STS task in Section 4.2.

4.1 Standard STS

4.1.1 Setups

Dataset and Metric We use seven standard STS datasets including STS tasks 2012-2016 (Agirre et al., 2012, 2013, 2014, 2015, 2016), STS Benchmark (Cer et al., 2017) and SICK-Relatedness (Marelli et al., 2014) for our experiments. Texts of these datasets are from news, forums, lexical definitions, etc. Each sample in these datasets contains a pair of sentences as well as a semantic similarity score ranging from 0 to 5. We use SentEval toolkit (Conneau and Kiela, 2018) for evaluation and report the Spearman's correlation on test sets following previous works (Reimers and Gurevych, 2019; Gao et al., 2021).

Baselines We compare unsupervised and supervised PromCSE to previous state-of-the-art sentence embedding methods. Unsupervised baselines comprise average GloVe embeddings (Pennington et al., 2014), average BERT embeddings (Gao et al., 2021), and post-processing methods such as BERT-flow (Li et al., 2020) and BERT-whitening (Su et al., 2021). We also introduce strong unsupervised baselines using contrastive learning, including IS-BERT (Zhang et al., 2020), CT-BERT (Carlsson et al., 2020), ConSERT (Yan et al., 2021), Mirror-BERT (Liu et al., 2021a), SG-OPT (Kim et al., 2021), SimCSE (Gao et al., 2021), DiffCSE (Chuang et al., 2022) and PromptBERT (Jiang et al., 2022). Methods taking extra supervision include InferSent (Conneau et al., 2017), Universal Sentence Encoder (Cer et al., 2018), SBERT (Reimers and Gurevych, 2019) along with applying BERT-flow, whitening and CT on it, ConSERT (Yan et al., 2021) and SimCSE (Gao et al., 2021).

Implementation Details We implement our models based on Huggingface's transformers (Wolf et al., 2020), where we also obtain the pre-trained checkpoints of BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019). We use the identical training data as SimCSE (Gao et al., 2021). Specifically, we train unsupervised PromCSE on

Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
Unsupervised models								
GloVe embeddings (avg.)$^{\clubsuit}$	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
BERT$_{base}$ (first-last avg.)$^{\diamondsuit}$	39.70	59.38	49.67	66.03	66.19	53.87	62.06	56.70
BERT$_{base}$-flow$^{\diamondsuit}$	58.40	67.10	60.85	75.16	71.22	68.66	64.47	66.55
BERT$_{base}$-whitening$^{\diamondsuit}$	57.83	66.90	60.9	75.08	71.31	68.24	63.73	66.28
IS-BERT$_{base}$$^{\heartsuit}$	56.77	69.24	61.21	75.23	70.16	69.21	64.25	66.58
CT-BERT$_{base}$$^{\diamondsuit}$	61.63	76.80	68.47	77.50	76.48	74.31	69.19	72.05
ConSERT$_{base}$$^{\spadesuit}$	64.64	78.49	69.07	79.72	75.95	73.97	67.31	72.74
Mirror-BERT$_{base}$†	67.40	79.60	71.30	81.40	74.30	76.40	70.30	74.40
SG-OPT-BERT$_{base}$‡	66.84	80.13	71.23	81.56	77.17	77.23	68.16	74.62
SimCSE-BERT$_{base}$$^{\diamondsuit}$	68.40	82.41	74.38	80.91	78.56	76.85	**72.23**	76.25
DiffCSE-BERT$_{base}$$^{\blacklozenge}$	72.28	84.43	76.47	83.90	80.54	80.59	71.23	78.49
PromptBERT$_{base}$$^{\triangle}$	71.56	84.58	**76.98**	**84.47**	**80.60**	**81.60**	69.87	**78.54**
$*$ PromCSE-BERT$_{base}$	**73.03**	**85.18**	76.70	84.19	79.69	80.62	70.00	78.49
Supervised models								
InferSent-GloVe$^{\clubsuit}$	52.86	66.75	62.15	72.77	66.87	68.03	65.65	65.01
Universal Sentence Encoder$^{\clubsuit}$	64.49	67.80	64.61	76.83	73.18	74.92	76.69	71.22
SBERT$_{base}$$^{\clubsuit}$	70.97	76.53	73.19	79.09	74.30	77.03	72.91	74.89
SBERT$_{base}$-flow$^{\diamondsuit}$	69.78	77.27	74.35	82.01	77.46	79.12	76.21	76.60
SBERT$_{base}$-whitening$^{\diamondsuit}$	69.65	77.57	74.66	82.27	78.39	79.52	76.91	77.00
CT-SBERT$_{base}$$^{\diamondsuit}$	74.84	83.20	78.07	83.84	77.93	81.46	76.42	79.39
ConSERT-BERT$_{base}$$^{\spadesuit}$	74.07	83.93	77.05	83.66	78.76	81.36	76.77	79.37
SimCSE-BERT$_{base}$$^{\diamondsuit}$	75.30	84.67	80.19	85.40	80.82	84.25	80.39	81.57
$*$ SimCSE-BERT$_{base}$ (reproduce)	75.13	84.35	80.26	85.45	80.83	84.29	80.39	81.53
$*$ SimCSE-BERT$_{base}$ + EH	75.22	84.93	**81.37**	85.94	80.94	**84.78**	80.38	81.94
$*$ PromCSE-BERT$_{base}$	75.58	84.33	79.67	85.79	81.24	84.25	80.79	81.81
$*$ PromCSE-BERT$_{base}$ + EH	**75.96**	**84.99**	80.44	**86.83**	**81.30**	84.40	**80.96**	**82.13**
SimCSE-RoBERTa$_{base}$$^{\diamondsuit}$	76.53	85.21	80.95	86.03	82.57	85.83	80.50	82.52
$*$ SimCSE-RoBERTa$_{base}$ + EH	76.83	85.67	81.57	86.35	82.72	86.84	**80.56**	82.86
$*$ PromCSE-RoBERTa$_{base}$	76.75	85.86	80.98	86.51	83.51	**86.58**	80.41	82.94
$*$ PromCSE-RoBERTa$_{base}$ + EH	**77.51**	**86.15**	**81.59**	**86.92**	**83.81**	86.35	80.49	**83.26**
SimCSE-RoBERTa$_{large}$$^{\diamondsuit}$	77.46	87.27	82.36	86.66	83.93	86.70	81.95	83.76
$*$ SimCSE-RoBERTa$_{large}$ + EH	78.01	87.65	82.55	87.21	84.19	86.95	82.03	84.08
$*$ PromCSE-RoBERTa$_{large}$	79.14	88.64	83.73	87.33	84.57	87.84	82.07	84.76
$*$ PromCSE-RoBERTa$_{large}$ + EH	**79.56**	**88.97**	**83.81**	**88.08**	**84.96**	**87.87**	**82.43**	**85.10**

Table 1: The performance of different sentence embedding models on test sets of STS tasks (Spearman's correlation). The best performance and the second-best performance methods are denoted in bold and underlined fonts respectively. $\clubsuit$: results from (Reimers and Gurevych, 2019); $\diamondsuit$: results from (Gao et al., 2021); $\heartsuit$: results from (Zhang et al., 2020); $\spadesuit$: results from (Yan et al., 2021); $\dagger$: results from (Liu et al., 2021a); $\ddagger$: results from (Kim et al., 2021); $\blacklozenge$: results from (Chuang et al., 2022); $\triangle$: results from (Jiang et al., 2022); $*$: results from our experiments; + EH: adding the Energy-based Hinge loss as shown in Equation (12).

1 million randomly sampled sentences from English Wikipedia for one epoch, and train supervised PromCSE on the combination of MNLI (Williams et al., 2018) and SNLI (Bowman et al., 2015) datasets for ten epochs. The training proceeds with the default random seed 42 for one run, the same as SimCSE. The training details of hyperparameters are shown in Appendix A.

4.1.2 Results

Table 1 shows that our unsupervised PromCSE-BERT$_{base}$ significantly outperforms SimCSE-BERT$_{base}$ and raises the averaged Spearman's correlation from 76.25% to 78.49%. Besides, it can acquire competitive results with current state-of-the-art DiffCSE-BERT$_{base}$ and PromptBERT$_{base}$. Note that although PromptBERT applies prompting to contrastive learning, it requires fine-tuning the whole PLM and manually designing discrete prompts (Jiang et al., 2022). Using supervised NLI datasets, PromCSE also surpasses SimCSE consistently based on various PLMs. Incorporating the Energy-based Hinge loss under supervised settings can further enhance SimCSE as well as PromCSE consistently over multiple pre-trained backbone models. It pushes state-of-the-art results to 82.13% using BERT_{base} and 85.10% using RoBERTa_{large}.

4.2 Domain-Shifted STS

4.2.1 Setups

Dataset and Metric The cumbersome data annotation leads to few datasets for STS tasks. Fortunately, we find a dataset with a different domain from the training corpus and the standard STS tasks. Crisscrossed Captions (CxC) (Parekh et al., 2021) extends the English MS-COCO (Lin et al., 2014) 5k dev and test sets with continuous (0-5) human similarity annotations, and it supports evaluation for correlation measures that compare model rankings with rankings derived from human similarity judgments for text-text comparisons. We use the STS task of CxC, whose texts are all image captions, to evaluate the domain-shifted robustness of various sentence embedding models.

Due to CxC's dense annotation where the scores between many pairs are themselves correlated, we choose a sampled Spearman's bootstrap correlation as the evaluation metric following (Parekh et al., 2021). For each correlation estimate, we sample half of the queries and for each selected query, we choose one of the items for which CxC supplies a paired rating. We compute Spearman's r between the CxC scores and the model scores for the selected pairs. The final correlation is the average over 1000 of these bootstrap samples.

Baselines We compare our unsupervised and supervised models to current SOTA sentence embedding methods. Unsupervised baselines include average GloVe embeddings (Pennington et al., 2014), SimCSE (Gao et al., 2021), DiffCSE (Chuang et al., 2022) and PromptBERT (Jiang et al., 2022). We choose SimCSE (Gao et al., 2021) as the supervised baseline. For reference, we also report two strong baselines ALIGN (Jia et al., 2021) and MURAL (Jain et al., 2021), which are trained specifically on MS-COCO.

4.2.2 Results

Table 2 demonstrates that by directly testing model checkpoints on the domain-shift CxC-STS dataset without further training, our unsupervised PromCSE remarkably boosts the performance of SimCSE by 3.7%, with a much more significant gap than 2.2% on standard STS tasks. Unsupervised PromCSE even outperforms state-of-the-art DiffCSE and PromptBERT by 1.1% and 1.2%, respectively. Compared with supervised SimCSE, PromCSE also achieves greater improvements on the CxC-STS task than on standard STS tasks, indicating better resilience to domain shifts. It is remarkable that our supervised PromCSE + EH could even *outperform* ALIGN and MUTUAL that are trained with in-domain MS-COCO annotations, reaching new state-of-the-art results.

Model	**CxC-STS** avg ± std
GloVe embeddings (avg.)♣	55.1 ± 0.6
∗ unsup-SimCSE-BERT_{base}	67.5 ± 1.2
∗ unsup-DiffCSE-BERT_{base}	70.1 ± 1.1
∗ unsup-PromptBERT$_{base}$	70.0 ± 1.1
∗ unsup-PromCSE-BERT_{base}	**71.2 ± 1.1**
∗ sup-SimCSE-BERT_{base}	73.0 ± 1.1
∗ sup-SimCSE-BERT_{base} + EH	73.2 ± 1.0
∗ sup-PromCSE-BERT_{base}	73.6 ± 1.0
∗ sup-PromCSE-BERT_{base} + EH	**74.0 ± 1.0**
ALIGN-BERT_{base}♣	72.7 ± 0.4
MURAL-BERT_{base}♣	73.9 ± 0.4

Table 2: Spearman's R Bootstrap Correlation (×100) on MS-COCO 5k test set using CxC annotations. ♣: results from (Jain et al., 2021); ∗ : results from our experiments.

Model	**Avg. STS**	**CxC-STS**
SimCSE	76.25	67.5
PromCSE	**78.49**	**71.2**
layer-shared soft prompt	77.64	71.0
input-layer soft prompt	68.35	67.4

Table 3: Test results of seven standard STS tasks (Avg. STS) and the CxC-STS task under different prompt types.

5 Ablation Studies

We investigate how different ways of choosing prompt type, prompt length and margin m affect our models. We use BERT_{base} model to evaluate on seven standard STS tasks and the CxC-STS task.

Type of Soft Prompt In PromCSE, we prepend multi-layer soft prompts to PLMs instead of only the input (embedding) layer as (Lester et al., 2021). Table 3 shows that only prepending soft prompts to the input layer significantly jeopardizes the performance of PromCSE on both standard STS tasks and the CxC-STS task. While making the weights of soft prompts shared across layers does not influence the effectiveness much.

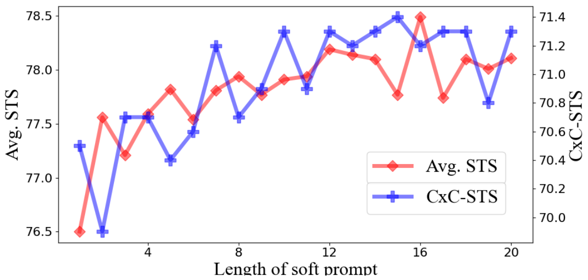


Figure 2: Test results of seven standard STS tasks (Avg. STS) and the CxC-STS task under various lengths of soft prompts.

m	*w/o*	*0*	*0.05*	*0.1*	*0.15*
Avg. STS	81.53	81.56	81.73	81.75	81.87
m	*0.2*	*0.25*	*0.3*	*0.35*	*0.4*
Avg. STS	**81.94**	81.91	81.71	81.48	81.36

Table 4: The average test set results of seven standard STS tasks under different margin m.

Length of Soft Prompt The soft prompts in PromCSE consist of a sequence of k trainable vectors. Here we regard k as the length of soft prompts and investigate its effect. Figure 2 shows that the model performance on standard STS tasks and the CxC-STS task rises as we increase the length of soft prompts, and finally tends to stabilize when k reaches around 12. It is interesting to observe that even with k set to 1, our PromCSE can still outperform SimCSE by 0.25% on standard STS tasks and 3% on the CxC-STS task, which indicates the effectiveness and robustness of our method.

Margin m The margin m in Energy-based Hinge loss (Equation (11)) controls the strength of the pairwise discriminative power. As shown in Table 4, the best performance is achieved when $m = 0.2$, either larger or smaller margin degrade the performance. This matches our intuition that small m may have little effect, and large m may overextend the distance between negative pairs.

6 Alignment and Uniformity Analysis

Alignment and uniformity are two properties proposed by (Wang and Isola, 2020) to measure the quality of representations. Specifically, given the distribution of positive pairs p_{pos} and the distribution of the whole dataset p_{data}, *alignment* computes the expected distance between normalized embeddings of the paired sentences:

$$\ell_{align} \triangleq \mathop{\mathbb{E}}_{(x,x^+)\backsim p_{pos}} \parallel f(x) - f(x^+) \parallel^2 \quad (13)$$

Model	Align	Uniform
BERT$_{base}$ (first-last avg.)	0.195	-1.304
unsup-SimCSE-BERT$_{base}$	0.238	**-2.337**
unsup-PromCSE-BERT$_{base}$	**0.117**	-1.354
sup-SimCSE-BERT$_{base}$	**0.241**	-3.246
sup-SimCSE-BERT$_{base}$ + EH	0.260	-3.349
sup-PromCSE-BERT$_{base}$	0.325	-3.268
sup-PromCSE-BERT$_{base}$ + EH	0.366	**-3.397**

Table 5: Alignment and Uniformity measured on STS-B. The smaller numbers are better.

While *uniformity* measures how well the embeddings are uniformly distributed in the representation space:

$$\ell_{uniform} \triangleq \log \mathop{\mathbb{E}}_{x,y \overset{i.i.d.}{\backsim} p_{data}} e^{-2\|f(x)-f(y)\|^2} \quad (14)$$

It can be seen in Table 5 that unsupervised PromCSE and supervised PromCSE are optimizing the representation space in two different directions. Compared with SimCSE, unsupervised PromCSE acquires better alignment, while supervised PromCSE has better uniformity. Besides, the Energy-based Hinge loss improves the uniformity of supervised models, which verifies its effectiveness in enhancing the pairwise discriminative power. To directly look into the representation space of different models, we visualize the cosine similarity distribution of sentence pairs from STS-B dataset for both SimCSE and PromCSE in Appendix B. It can be observed in Figure 3 that unsupervised PromCSE preserves a lower variance while supervised PromCSE shows a more scattered distribution compared to SimCSE, corresponding to better alignment and uniformity, respectively.

7 Conclusion

This paper presents PromCSE, a prompt-based contrastive learning framework that improves universal sentence embeddings for resilience to domain shifts. Additionally, we theoretically show that the contrastive learning framework under NT-Xent loss is an instance of energy-based learning. To further boost the performance of supervised sentence embeddings, we propose an Energy-based Hinge loss to supplement NT-Xent loss. Extensive experiments on seven STS tasks and one domain shift STS task both verify the effectiveness of our method compared to current state-of-the-art supervised and unsupervised sentence embedding models.

Limitations

In this section, we illustrate the limitations of our method. Firstly, although PromCSE outperforms SimCSE on STS tasks under both unsupervised and supervised settings, it cannot boost the performance of SimCSE on supervised transfer tasks, as shown in Appendix C. We share a similar sentiment with (Reimers and Gurevych, 2019) that the primary goal of sentence embeddings is to cluster semantically similar sentences. Hence, we take STS results as the main comparison in this paper. Secondly, our proposed Energy-based Hinge loss is shown to be useful when hard negatives are available in supervised NLI datasets. However, how to automatically sample or generate hard negatives with unlabeled data is not discussed in this paper. We believe that designing algorithms that can automatically retrieve hard negatives will be a good direction for future work to improve the performance of unsupervised sentence embeddings.

Ethics Statement

Since our method relies on pre-trained language models, it may run the danger of inheriting and propagating some of the models' negative biases from the data they have been pre-trained on (Bender et al., 2021). Furthermore, we do not see any other potential risks.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel M. Cer, Mona T. 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@NAACL-HLT 2015, Denver, Colorado, USA, June 4-5, 2015*, pages 252–263. The Association for Computer Linguistics.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel M. Cer, Mona T. 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@COLING 2014, Dublin, Ireland, August 23-24, 2014*, pages 81–91. The Association for Computer Linguistics.

Eneko Agirre, Carmen Banea, Daniel M. Cer, Mona T. 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@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, pages 497–511. The Association for Computer Linguistics.

Eneko Agirre, Daniel M. Cer, Mona T. Diab, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the 6th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2012, Montréal, Canada, June 7-8, 2012*, pages 385–393. The Association for Computer Linguistics.

Eneko Agirre, Daniel M. Cer, Mona T. Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *sem 2013 shared task: Semantic textual similarity. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics, *SEM 2013, June 13-14, 2013, Atlanta, Georgia, USA*, pages 32–43. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, pages 610–623. ACM.

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, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 632–642. The 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. 2020. Semantic re-tuning with contrastive tension. In *International Conference on Learning Representations*.

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, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 169–174. Association for Computational Linguistics.

Daniel M. Cer, Mona T. 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@ACL 2017, Vancouver, Canada, August 3-4, 2017*, pages 1–14. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020a. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020b. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljačić, Shang-Wen Li, Wen-tau Yih, Yoon Kim, and James Glass. 2022. Diffcse: Difference-based contrastive learning for sentence embeddings. *arXiv preprint arXiv:2204.10298*.

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).

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, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 670–680. Association for Computational Linguistics.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4690–4699. Computer Vision Foundation / IEEE.

Yuntian Deng, Anton Bakhtin, Myle Ott, Arthur Szlam, and Marc'Aurelio Ranzato. 2020. Residual energy-based models for text generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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.

Bill Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Third International Workshop on Paraphrasing (IWP2005)*.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and GPT-2 embeddings. 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 55–65. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021a. 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.

John M. Giorgi, Osvald Nitski, Bo Wang, and Gary D. Bader. 2021b. 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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 879–895. Association for Computational Linguistics.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 1367–1377. The Association for Computational Linguistics.

Geoffrey E. Hinton. 2002. Training products of experts by minimizing contrastive divergence. *Neural Comput.*, 14(8):1771–1800.

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*, pages 168–177.

Aashi Jain, Mandy Guo, Krishna Srinivasan, Ting Chen, Sneha Kudugunta, Chao Jia, Yinfei Yang, and Jason Baldridge. 2021. MURAL: Multimodal, multitask representations across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3449–3463, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR.

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. *CoRR*, abs/2201.04337.

Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. 2021. Self-guided contrastive learning for bert sentence representations. *arXiv preprint arXiv:2106.07345*.

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.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. 2006. A tutorial on energy-based learning. *Predicting structured data*, 1(0).

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. 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 3045–3059. Association for Computational Linguistics.

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 2020, Online, November 16-20, 2020*, pages 9119–9130. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021a. Fast, effective, and self-supervised: Transforming masked language models into universal lexical and sentence encoders. *arXiv preprint arXiv:2104.08027*.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021b. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *CoRR*, abs/2110.07602.

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.

Xiaofei Ma, Peng Xu, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2019. Domain adaptation with bert-based domain classification and data selection. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 76–83.

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 2014, Reykjavik, Iceland, May 26-31, 2014*, pages 216–223. European Language Resources Association (ELRA).

Andriy Mnih and Yee Whye Teh. 2012. A fast and simple algorithm for training neural probabilistic language models. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress.

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 (ACL-04)*, pages 271–278.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd*

Annual Meeting of the Association for Computational Linguistics (ACL'05), pages 115–124.

Zarana Parekh, Jason Baldridge, Daniel Cer, Austin Waters, and Yinfei Yang. 2021. Crisscrossed captions: Extended intramodal and intermodal semantic similarity judgments for MS-COCO. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2855–2870, Online. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language models as knowledge bases? 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 2463–2473. Association for Computational Linguistics.

Marc'Aurelio Ranzato, Y-Lan Boureau, Sumit Chopra, and Yann LeCun. 2007. A unified energy-based framework for unsupervised learning. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, AISTATS 2007, San Juan, Puerto Rico, March 21-24, 2007*, volume 2 of *JMLR Proceedings*, pages 371–379. JMLR.org.

Marc'Aurelio Ranzato, Christopher S. Poultney, Sumit Chopra, and Yann LeCun. 2006. Efficient learning of sparse representations with an energy-based model. In *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, pages 1137–1144. MIT Press.

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 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.

Timo Schick and Hinrich Schütze. 2020a. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*.

Timo Schick and Hinrich Schütze. 2020b. It's not just size that matters: Small language models are also few-shot learners. *arXiv preprint arXiv:2009.07118*.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 815–823. IEEE Computer Society.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.

Ziyi Shou, Yuxin Jiang, and Fangzhen Lin. 2022. AMR-DA: Data augmentation by Abstract Meaning Representation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3082–3098, Dublin, Ireland. Association for Computational Linguistics.

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*, pages 1631–1642.

Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. 2016. Deep metric learning via lifted structured feature embedding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4004–4012. IEEE Computer Society.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Feng Wang and Huaping Liu. 2021. Understanding the behaviour of contrastive loss. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 2495–2504. Computer Vision Foundation / IEEE.

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. 2018. Cosface: Large margin cosine loss for deep face recognition. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5265–5274. Computer Vision Foundation / IEEE Computer Society.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 9929–9939. PMLR.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2):165–210.

Adina Williams, Nikita Nangia, and Samuel R. 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, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. Consert: A contrastive framework for self-supervised sentence representation transfer. *arXiv preprint arXiv:2105.11741*.

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.

Yuhao Zhang, Hongji Zhu, Yongliang Wang, Nan Xu, Xiaobo Li, and Binqiang Zhao. 2022. A contrastive framework for learning sentence representations from pairwise and triple-wise perspective in angular space. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4892–4903, Dublin, Ireland. Association for Computational Linguistics.

Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual probing is [mask]: Learning vs. learning to recall. *arXiv preprint arXiv:2104.05240*.

Kun Zhou, Beichen Zhang, Xin Zhao, and Ji-Rong Wen. 2022. Debiased contrastive learning of unsupervised sentence representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6120–6130, Dublin, Ireland. Association for Computational Linguistics.

A Training Details

We conduct experiments on 4 NVIDIA 3090Ti GPUs. The maximum sequence length is set to 32, and the temperature τ in NT-Xent loss is set to 0.05. Adam optimizer is used with a linear decay schedule. We use grid-search of batch size $\in \{256, 512\}$, initial learning rate $\in$ {5e-3, 1e-2, 3e-2} (prompt tuning requires relative larger initial learning rate than fine-tuning) and prompt length $\in \{10, 12, 14, 16\}$. During the training process, we save the checkpoint with the highest score on the STS-B development set, by evaluating our model every 125 training steps. And then we use STS-B development set to find the best hyperparameters (listed in Table 6).

	Unsupervised	Supervised		
	BERT	BERT	RoBERTa	
	base	base	base	large
Batch size	256	256	512	512
Learning rate	3e-2	1e-2	1e-2	5e-3
Prompt length	16	12	10	10

Table 6: The main hyperparameters for PromCSE in standard STS tasks.

As for Energy-based Hinge loss, the margin m is set to 0.2 according to the ablation study in Section 5. When adding Energy-based Hinge loss to supervised SimCSE, we do not change the training configurations of the original SimCSE.

For both unsupervised and supervised PromCSE, we take the $[CLS]$ representation with an MLP layer on top of it as the sentence representation. Specially, for unsupervised PromCSE, we discard the MLP layer and only use the $[CLS]$ output during test, the same as SimCSE (Gao et al., 2021).

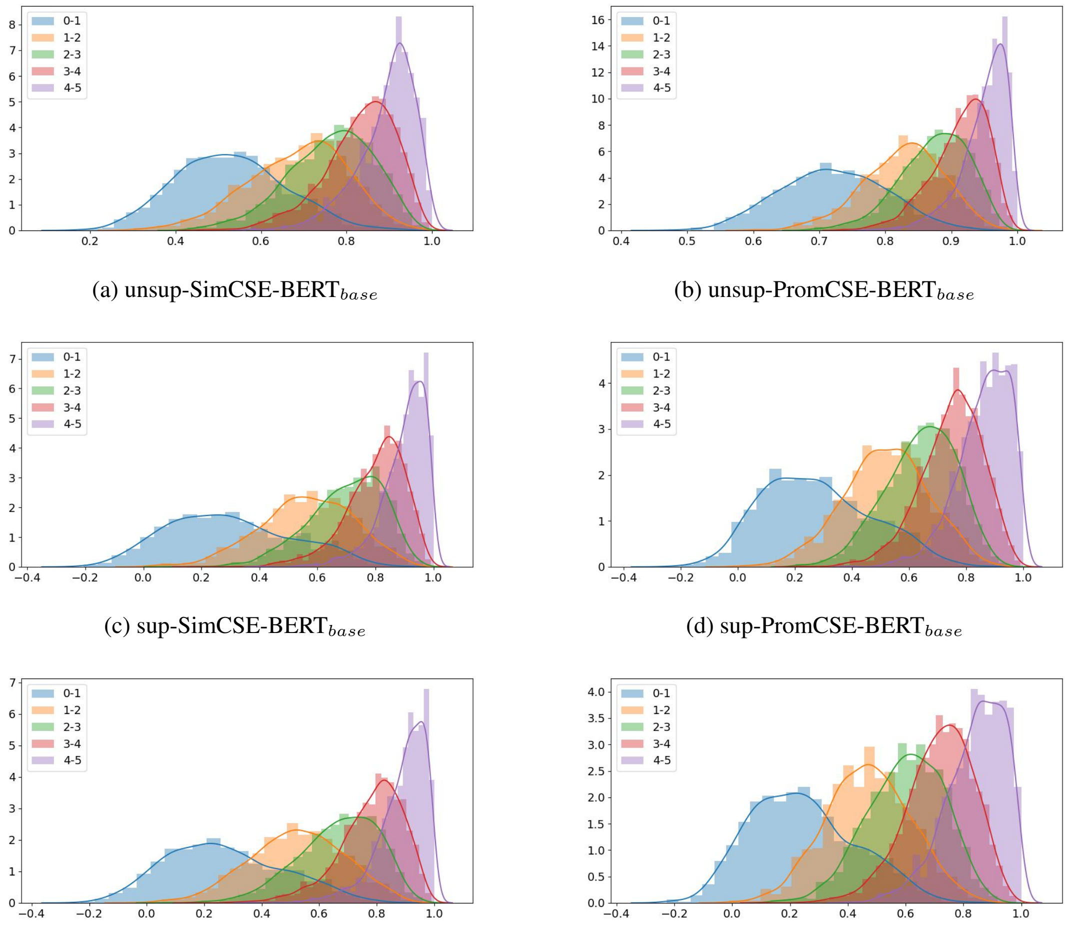
Prompt Initialization (Li and Liang, 2021) find that the parameter initialization of the *Soft Prompt* has a significant impact in low-data settings. Though our unsupervised and supervised training data both exceed 100,000, we still attempted various initialization strategies for soft prompts of PromCSE including (1) random initialization; (2) initializing with manual discrete prompt like *"The meaning of the sentence"*; (3) using an LSTM to generate the sequence of *Soft Prompt*; (4) first pretraining *Soft Prompt* by training PromCSE using the Masked Language Modeling (MLM) objective on the training data. However, we find that different initialization strategies do not have much impact on our tasks. As a result, we randomly initialize the soft prompts using the default $init\_weights$ function provided by Huggingface's transformers (Wolf et al., 2020) for all the experiments.

B Distribution of Sentence Embeddings

We visualize the cosine similarity density plots of various models on the STS-Benchmark dataset in Figure 3. Concretely, we split the STS-B dataset into five similarity levels according to their golden labels and count all similarity scores in each sentence level.

C Supervised Transfer Tasks for Sentence Embeddings

Following (Gao et al., 2021), we evaluate our models with SentEval toolkit (Conneau and Kiela, 2018) on several supervised transfer tasks, including: 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) and MRPC (Dolan and Brockett, 2005). A logistic regression classifier is trained on top of (frozen) sentence embeddings produced by different methods. The evaluation results are listed in Table 7 for reference.



(a) unsup-SimCSE-BERT$_{base}$

(b) unsup-PromCSE-BERT$_{base}$

(c) sup-SimCSE-BERT$_{base}$

(d) sup-PromCSE-BERT$_{base}$

(e) sup-SimCSE-BERT$_{base}$ + EH

(f) sup-PromCSE-BERT$_{base}$ + EH

Figure 3: Cosine Similarity Density Plots of different models between sentence pairs in STS-B. Pairs are divided into five groups based on ground truth ratings (higher means more similar). The x-axis is the model predicted cosine similarity.

Model	**MR**	**CR**	**SUBJ**	**MPQA**	**SST-2**	**TREC**	**MPRC**	**Avg.**
Unsupervised models								
GloVe embeddings (avg.)♣	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
BERT$_{base}$ (first-last avg.)♣	78.66	86.25	94.37	88.66	84.40	92.80	69.54	84.94
BERT$_{base}$ (CLS)♣	78.68	84.85	94.21	88.23	84.13	91.40	71.13	84.66
IS-BERT$_{base}$♡	81.09	**87.18**	**94.96**	88.75	**85.96**	88.64	74.24	**85.83**
SimCSE-BERT$_{base}$♢	**81.18**	86.46	94.45	88.88	85.50	**89.80**	74.43	85.81
$*$ PromCSE-BERT$_{base}$	80.95	85.46	94.50	**89.46**	84.84	88.40	**74.61**	85.46
Supervised models								
InferSent-GloVe♣	81.57	86.54	92.50	90.38	84.18	88.20	75.77	85.59
Universal Sentence Encoder♣	80.09	85.19	93.98	86.70	86.38	**93.20**	70.14	85.10
SBERT$_{base}$♣	**83.64**	**89.43**	94.39	89.86	**88.96**	89.60	**76.00**	**87.41**
SimCSE-BERT$_{base}$♢	82.69	89.25	**94.81**	89.59	87.31	88.40	73.51	86.51
$*$ SimCSE-BERT$_{base}$ + EH	82.81	88.82	94.34	89.98	88.14	86.20	74.90	86.46
$*$ PromCSE-BERT$_{base}$	81.86	88.56	93.78	89.69	86.44	82.80	75.36	85.50
$*$ PromCSE-BERT$_{base}$ + EH	81.80	89.85	93.92	**90.72**	87.05	82.60	75.43	85.91

Table 7: Transfer task results of different sentence embedding models (measured as accuracy). ♣: results from (Reimers and Gurevych, 2019); ♡: results from (Zhang et al., 2020); ♢: results from (Gao et al., 2021); $*$: results from our experiments; + EH: adding the Energy-based Hinge loss as shown in Equation (12).