LEARNING TO PERTURB FOR CONTRASTIVE LEARNING OF UNSUPERVISED SENTENCE REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently, contrastive learning has been shown effective in fine-tuning pre-trained language models (PLM) to learn sentence representations, which incorporates perturbations into unlabeled sentences to augment semantically related positive examples for training. However, previous works mostly adopt heuristic perturbation methods that are independent of the sentence representations. Since the perturbations are unaware of the goal or process of sentence representation learning during training, it is likely to lead to sub-optimal augmentations for conducting constrative learning. To address this issue, we propose a new framework **L2P-CSR** that adopts a learnable perturbation strategy for improving contrastive learning of sentence representations. In our L2P-CSR, we design a safer perturbation mechanism that only weakens the influence of tokens and features on the sentence representation, which avoids dramatically changing the semantics of the sentence representations. Besides, we devise a gradient-based algorithm to generate adaptive perturbations specially for the dynamically updated sentence representation during training. Such a way is more capable of augmenting high-quality examples that guide the sentence representation learning. Extensive experiments on diverse sentence-related tasks show that our approach outperforms competitive baselines.

1 INTRODUCTION

Unsupervised sentence representation learning is a fundamental problem in natural language processing (NLP) Hill et al. (2016); Le & Mikolov (2014), which aims to learn effective sentence representations that can benefit various downstream tasks. High-quality unsupervised sentence representations are essential to low-resource domains or computationally expensive NLP tasks Cer et al. (2017), including zero-shot semantic similarity comparison Agirre et al. (2016) and large-scale document retrieval Le & Mikolov (2014). Recently, BERT-based pre-trained language models (PLMs) Devlin et al. (2019) have achieved excellent performance on various NLP tasks. However, since these PLMs are mainly pre-trained with token-level self-supervised loss (*e.g.,* masked language model), the derived sentence representations may not be best suited to sentence-level tasks Li et al. (2020); Yan et al. (2021). Previous works find that the sentence representations from BERT-based models are not uniformly distributed with respect to direction but occupy a narrow cone in the vector space Ethayarajh (2019), which largely limits their semantic representation capacity.

To solve it, contrastive learning Chen et al. (2020); He et al. (2020) has been proposed to fine-tune PLMs for refining the sentence representations. The basic idea is to pull semantically close neighbors together while pushing apart non-neighbors. Concretely, given a set of unlabeled texts, existing approaches first augment different but semantically related examples by incorporating *perturbations* (*e.g.,* token shuffling Yan et al. (2021) and dropout Gao et al. (2021)) into the original sentence. Then the augmented data will be used as the positive example pairs to fine-tune PLMs to learn the contrastive objective. In this way, both the alignment between semantically related positive pairs and the uniformity of the whole representation space can be substantially improved.

To effectively conduct contrastive learning, it is key to design a proper perturbation strategy to augment positive examples. However, existing works mostly rely on heuristic perturbation methods. Despite the performance improvement, there are two major shortcomings that are likely to affect the sentence representation learning. First, heuristic methods are usually unaware of the goal or process of sentence representation learning, which can't produce pertinent perturbations for adaptively

improving the sentence representations. As a result, these augmented samples are likely to fall in a constrained semantic space limited by the heuristic strategies, leading to sub-optimal performance or even overfitting. Second, since PLMs are often over-parameterized Houlsby et al. (2019), the derived sentence representations are sensitive to the perturbations in input Gao et al. (2021). Small perturbations *e.g.,* dropout, may result in significantly different representations. However, radical heuristic perturbations (*e.g.,* token cutoff and shuffling) have been widely used in previous works. Although leading to substantial differences compared with original inputs, it may hurt the semantics of the augmented examples and mislead the learning of the contrastive objective.

To address the above issues, we propose a new contrastive learning framework with a learnable perturbation strategy for unsupervised sentence representation learning. The core idea is learning to generate more informative perturbations that are adaptive to dynamically changing sentence representations during training and meanwhile guarantee the semantic consistency in the representation space. Specifically, we introduce a safer perturbation mechanism by using weakening masks to halve the representations of perturbed tokens and features, and devise a gradient-based algorithm to optimize the perturbations towards augmenting the most confusing views for sentence representation learning. Compared with heuristic perturbations in previous works Yan et al. (2021), our perturbation mechanism mainly focuses on *weakening* but not *removing* tokens or features, which avoids large changes in the semantics of sentence representations. In addition, the gradient-based perturbation algorithm can generate informative augmentations focusing on the most confusing views in sentence representation learning during training. In this way, the perturbation approach can adaptively augment the positive examples specially for the intermediate sentence representations that are not well learned, which gradually improves the sentence representation capacity of PLMs.

To this end, we propose **L2P-CSR**, a general framework to Learn how to Perturb for Contrastive learning of unsupervised Sentence Representations. In L2P-CSR, we first adopt randomly initialized probability matrices to generate the token-level and feature-level weakening masks to perturb the PLM. Then, we leverage a gradient-based algorithm to update the perturbations by considering the intermediate sentence representations. Finally, based on the perturbations, we augment two different representations of the same sentence for contrastive learning. We demonstrate that our L2P-CSR outperforms competitive baselines on semantic textual similarity tasks and transfer tasks.

To our knowledge, our approach is the first attempt to explore learning to perturb for contrastive learning of unsupervised sentence representations, which can adaptively generate informative augmentations during the training of sentence representation models. Extensive experiments have demonstrated the effectiveness of the proposed approach against a number of competitive baselines.

2 Related Work

Sentence Representation Learning. Existing works of sentence representation learning can be categorized into supervised Cer et al. (2018); Gao et al. (2021) and unsupervised approaches Hill et al. (2016); Zhang et al. (2022). Supervised approaches rely on annotated datasets to train the encoder network Cer et al. (2018); Zhang et al. (2021a) for producing sentence representations, *e.g.,* NLI datasets Williams et al. (2018). Unsupervised approaches consider deriving sentence representations without labeled datasets. Early works find that simply pooling word embeddings Pennington et al. (2014) leads to strong results. Recently, pre-trained language models Devlin et al. (2019) have shown effectiveness in NLP tasks but their produced representations suffer from the anisotropy problem Giorgi et al. (2021); Wu et al. (2020). Several works propose to regularize the representations, *e.g.,* flow-based approach Li et al. (2020) and whitening method Huang et al. (2021). Besides, recent works Wang et al. (2021); Liu et al. (2021); Ni et al. (2021) adopt contrastive objectives with data augmentation methods on unsupervised datasets to refine the representations of PLMs.

Contrastive Learning. Contrastive learning has been a popular technique in compute vision area with solid performance He et al. (2020); Chen et al. (2020). Usually, it requires a set of positive examples that are semantically related. A surge of works apply data augmentation strategies on unlabeled data to augment positive examples, *e.g.,* random cropping and rotation Chen et al. (2020); Yan et al. (2021). For sentence representation learning, contrastive learning can achieve a better alignment-uniformity balance. Several works adopt back translation Fang & Xie (2020), token shuffle and cutoff Yan et al. (2021) to augment positive examples for representations learning.

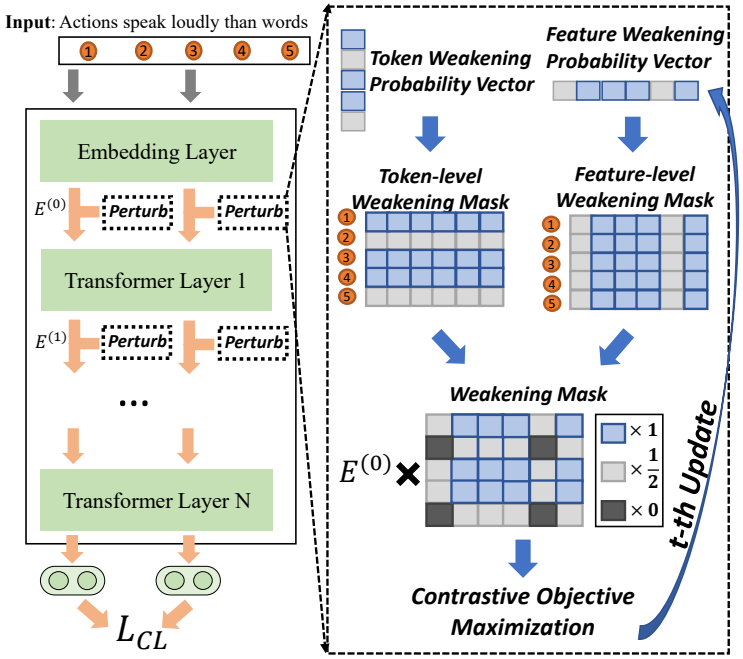


Figure 1: The L2P-CSR framework. We show the case that perturbs the representations of the first two layers in the t-th iterative update.

Recently, SimCSE Gao et al. (2021) uses Dropout acts for augmentation, and performs well on STS tasks. Besides, instead of data augmentation, SG-OPT Kim et al. (2021) utilizes the self-guidance mechanism and VaSCL Zhang et al. (2021b) presents a virtual augmentation-oriented framework for contrastive learning of unsupervised sentence representations.

3 Preliminary

This work seeks to make use of unlabeled corpus to learn effective sentence representations that can be directly utilized in downstream tasks, *e.g.,* STS task Agirre et al. (2016). Given a set of unlabeled sentences $\mathcal{X} = \{x_1, \ldots, x_n\}$, our goal is to learn a representation $\mathbf{h}_i \in \mathcal{R}^d$ for each sentence x_i in an unsupervised manner. We denote this process with a parameterized function $\mathbf{h}_i = f(x_i)$.

Here, we mainly focus on using BERT-based PLMs to generate sentence representations. For each sentence consisting of a sequence of tokens as $x_i = \{w_1, w_2, \ldots, w_m\}$, PLMs first project these tokens into a token embedding matrix $\mathbf{E}^{(0)} \in \mathcal{R}^{m \times d}$ via the embedding layer, and then a stack of Transformer layers will gradually encode it to generate the l-th layer representations $\mathbf{E}^{(l)} \in \mathcal{R}^{m \times d}$. Following existing works Li et al. (2020), we fine-tune the PLMs on the unlabeled corpus via our proposed unsupervised learning method. For each sentence x_i, we encode it by the fine-tuned PLMs and take the representation of the `[CLS]` token from the last layer as its representation $\mathbf{h}_i$.

4 Approach

Our proposed framework L2P-CSR is to improve the perturbation strategy of the contrastive learning paradigm for sentence representation learning. In L2P-CSR, we devise a learnable perturbation strategy containing a safer perturbation mechanism for semantic consistency and a gradient-based algorithm to adapt the perturbations to the dynamically changing sentence representations during training. Concretely, we design the token-level and feature-level weakening masks to halve part of values from the representations of the first few layers as perturbations, and the weakening masks can be iteratively updated by contrastive objective maximization. Then, we utilize the perturbations to augment representations for contrastive learning. The overview of our L2P-CSR is shown in Fig. 1.

4.1 PERTURBATION WITH WEAKENING MASKS

We aim to develop a perturbation mechanism that can not radically change the sentence semantics but also introduces differences. Given the input sentence x_i, we design the token-level and feature-level perturbation masks to weaken its representations in the first k layers $\{\mathbf{E}^{(l)}\}_{l=0}^{k-1}$. [1]

For the representations of each layer, we construct different token-level and feature-level masks. To generate the token-level mask of the l-th layer, we first obtain a token-weakening probability vector $\mathcal{P}_{tok}^{(l)} = \{p_1^t, p_2^t, \ldots, p_m^t\}$, where $p_i^t \in [0,1]$ denotes the probability whether the representation of the i-th token should not be weakened. For feature-level mask, we also acquire a probability vector $\mathcal{P}_{fea}^{(l)} = \{p_1^f, p_2^f, \ldots, p_d^f\}$ for the l-th layer, where $p_j^f \in [0,1]$ denotes the probability whether the j-th feature should not be weakened. Based on the probability vectors, we can obtain the token-level mask $\mathbf{M}_t = \{\alpha_1^t, \alpha_2^t, \ldots \alpha_m^t\}$ and feature-level mask $\mathbf{M}_f = \{\alpha_1^f, \alpha_2^f, \ldots \alpha_d^f\}$ as

$$\alpha_i = \begin{cases} 1, \ p_i \geq \phi \\ 0, \ p_i < \phi \end{cases}, \tag{1}$$

where ϕ is a hyperparameter of the mask threshold, α_i denotes the mask value of token (α_i^t) or feature (α_i^f), p_i denotes the weakening probability of the i-th token (p_i^t) or feature (p_i^f). Then, we average the token-level and feature-level weakening masks to produce the weakening mask $\mathbf{M}^{(l)}$ for the output representations of the l-th layer $\mathbf{E}^{(l)}$. Specifically, for the j-th feature of the i-th token, its corresponding value α_{ij} of the weakening mask $\mathbf{M}^{(l)}$ is

$$\alpha_{ij} = (\alpha_i^t + \alpha_j^f)/2. \tag{2}$$

Finally, we incorporate the weakening mask as perturbations on the first k layers. For the l-th layer, we directly multiply its output representations $\mathbf{E}^{(l)}$ and the corresponding weakening mask $\mathbf{M}^{(l)}$ as

$$\tilde{\mathbf{E}}^{(l)} = \mathbf{E}^{(l)} \times \mathbf{M}^{(l)}. \tag{3}$$

In this way, if a token or feature is selected to be weakened, its corresponding token representation or feature values will be halved (*i.e.,* $\times \frac{1}{2}$). Only if the token and the feature are both weakened, the value will be zero. Compared with existing works using token cutoff or reordering Yan et al. (2021), our approach conducts a weaker perturbation with more controllable variations on the sentence representations. As a result, the representations from the last layer will be perturbed in a safer way, which alleviates the information loss caused by perturbation but also introduce difference. For each sentence x_i, we generate two perturbed representations $\tilde{\mathbf{h}}_i$ and $\tilde{\mathbf{h}}_i^+$ by applying different weakening masks using the above approach. The two representations will be utilized for contrastive learning.

4.2 LEARNING TO PERTURB FOR CONTRASTIVE LEARNING

Based on the above perturbation mechanism, we propose an algorithm for learning how to generate the weakening masks for effective contrastive learning. We first initialize the weakening probability vectors and then optimize these vectors to adapt to the intermediate sentence representations in each step. Finally, we leverage these probability vectors to produce the weakening masks as perturbations for contrastive learning.

For the output representations of the l-th layer, we first initialize the token-level and feature-level weakening probability vectors from the uniform distribution as

$$\mathcal{P}_{tok}^{(l)}, \mathcal{P}_{fea}^{(l)} \sim U(0,1). \tag{4}$$

The weakening probability vectors are then utilized to generate the weakening masks $\mathbf{M}^{(l)}$ for contrastive learning using Eq 1 and Eq 2. To make the weakening masks adapt to the dynamically changing sentence representations during training, we optimize the weakening masks with the consideration of the intermediate sentence representations. Inspired by virtual adversarial training Miyato et al. (2017); Zhu et al. (2020), we design a contrastive loss maximization objective to produce

[1] For simplicity, we denote the output token embedding matrix from the embedding layer as the representation of the 0-th layer $\mathbf{E}^{(0)}$.

gradients for perturbation update. However, the values of the weakening masks are discrete, which cannot be directly optimized by gradient-based algorithms. Therefore, we propose to leverage the gradients of the discrete weakening masks $\mathbf{M}_t^{(l)}$ and $\mathbf{M}_f^{(l)}$ to update the continuous-valued weakening probability vectors $\mathcal{P}_{tok}^{(l)}$ and $\mathcal{P}_{fea}^{(l)}$ as

$$\mathcal{P}_{tok}^{(l)} = \Pi_{\mathcal{P}_{tok}^{(l)} \in [0,1]}(\mathcal{P}_{tok}^{(l)} + \beta g(\mathbf{M}_t^{(l)})/||g(\mathbf{M}_t^{(l)})||_2), \tag{5}$$

$$\mathcal{P}_{fea}^{(l)} = \Pi_{\mathcal{P}_{fea}^{(l)} \in [0,1]}(\mathcal{P}_{fea}^{(l)} + \beta g(\mathbf{M}_f^{(l)})/||g(\mathbf{M}_f^{(l)})||_2), \tag{6}$$

where we constrain $\mathcal{P}_{tok}^{(l)}$ and $\mathcal{P}_{fea}^{(l)}$ within $[0, 1]$ since they reflect the probabilities, β is the learning rate, $||\cdot||_2$ is the $L2$-norm, and $g(\mathbf{M}_t^{(l)})$ and $g(\mathbf{M}_f^{(l)})$ are the gradients of $\mathbf{M}_t^{(l)}$ and $\mathbf{M}_f^{(l)}$ by maximizing the contrastive loss between the two perturbed representations of the same sentence as

$$g(\mathbf{M}_t^{(l)}) = \bigtriangledown_{\mathbf{M}_t^{(l)}} L_{CL}(\tilde{\mathbf{h}}_i, \tilde{\mathbf{h}}_i^+), \tag{7}$$

$$g(\mathbf{M}_f^{(l)}) = \bigtriangledown_{\mathbf{M}_f^{(l)}} L_{CL}(\tilde{\mathbf{h}}_i, \tilde{\mathbf{h}}_i^+). \tag{8}$$

In this way, the weakening probability vectors can learn to provide the most confusing weakening masks for augmenting perturbed representations $\tilde{\mathbf{h}}_i$ and $\tilde{\mathbf{h}}_i^+$. The two representations reflect two most different views of the sentence representations. Finally, we adopt the cross-entropy contrastive learning objective with in-batch negatives Chen et al. (2020) to learn the sentence representations as

$$L_{CL}(\tilde{\mathbf{h}}_i, \tilde{\mathbf{h}}_i^+) = \log \frac{\exp(\text{sim}(\tilde{\mathbf{h}}_i, \tilde{\mathbf{h}}_i^+)/\tau)}{\sum_{\tilde{\mathbf{h}}_i^- \in \{\tilde{\mathbf{h}}^-\}} \exp(\text{sim}(\tilde{\mathbf{h}}_i, \tilde{\mathbf{h}}_i^-)/\tau)}, \tag{9}$$

where $\{\tilde{\mathbf{h}}^-\}$ is the negative example set for positive examples $(\tilde{\mathbf{h}}_i, \tilde{\mathbf{h}}_i^+)$, τ is a temperature hyperparameter and $\text{sim}(\tilde{\mathbf{h}}_i, \tilde{\mathbf{h}}_i^+)$ is the cosine similarity $\frac{\tilde{\mathbf{h}}_i^\top \tilde{\mathbf{h}}_i^+}{||\tilde{\mathbf{h}}_i|| \cdot ||\tilde{\mathbf{h}}_i^+||}$.

4.3 Overview of L2P-CSR

Our framework L2P-CSR contains two important stages. In the first stage, we learn the perturbation to adapt itself for the intermediate sentence representations. Concretely, we first initialize the token-level and feature-level weakening probability vectors, and generate the corresponding weakening masks to perturb the sentence representations. Then, we incorporate a gradient-based algorithm to adjust the weakening probability vectors by maximizing the contrastive objective between two perturbed representations of a sentence. In this process, the weakening masks are refreshed every time the probability vectors are updated. After t-th iteration, we can obtain the weakening masks as the perturbations that reflect the most different views of the same sentence. In the second step, we leverage the weakening masks to augment the perturbed representations for contrastive learning. We present the training algorithm in Supplementary Material.

Analysis. Existing works Yan et al. (2021); Gao et al. (2021) rely on heuristic strategy to augment different views for contrastive learning. For the sentence x_i, the learning objective is:

$$\min \frac{1}{n} \sum_{(\epsilon_1, \epsilon_2) \sim p_{pos}} \left[L_{CL}(f(x_i + \epsilon_1), f(x_i + \epsilon_2)) \right], \tag{10}$$

where n is the sampling number, ϵ_1 and ϵ_2 are perturbations to augment different-view representations for x_i. Both perturbations are sampled from the pre-defined perturbation distribution p_{pos} (*e.g.,* random token cutoff). Since p_{pos} is usually independent of sentence representations, the sampled perturbations are also unaware of the goal or process of sentence representation learning. As a result, the sampled perturbations can not adjust themselves to capture the semantic changes of intermediate representations during training, and the sentence representations are easy to overfit into the fixed perturbation distribution. In contrast, our approach proposes learnable perturbations which can adapt with the intermediate representations to provide the most different views of the same sentence. Such a process can be seen as a Min-Max game:

$$\min \frac{1}{n} \sum \left[L_{CL}(f(x_i + \epsilon_1), f(x_i + \epsilon_2)) \right], \tag{11}$$
$$\text{where}(\epsilon_1, \epsilon_2) \approx \arg\max_{\epsilon_1, \epsilon_2} \left\{ L_{CL}(f(x_i + \epsilon_1), f(x_i + \epsilon_2)) \right\}.$$

In this way, the intermediate sentence representations can be utilized to generate more informative perturbations, which help augment the most confusing positive examples specially for the intermediate sentence representations. By learning such examples, the sentence representations can be gradually improved and the overfitting risk can be reduced.

Discussion. Our work provides a framework that learns to perturb for contrastive learning of sentence representations. As analyzed above, our approach is equivalent to a Min-Max game that generates the most confusing perturbations for the intermediate sentence representations. To avoid the perturbations drastically changing the semantics, we also devise a safer perturbation mechanism that only halves the representations of perturbed tokens or feature. Compared with heuristic perturbation methods (*e.g.,* token cutoff and shuffling), the safer perturbation mechanism guarantees the semantic consistency of perturbed representations, and the learning to perturb strategy adaptively improves the informativeness of the perturbations. In this way, we are more capable of augmenting high-quality examples to improve the sentence representations.

Our L2P-CSR is similar to virtual adversarial training methods Zhu et al. (2020) which add gradient-based noise on the embedding layer for improving the smoothness. However, due to the over-parameter nature of PLMs Houlsby et al. (2019), the added noise is easy to hurt the semantics of the representations. Therefore, our L2P-CSR utilizes a safer perturbation strategy using the weakening masks, which rarely erase the input information, and only reduce the influence of several tokens and features on the sentence representation. In this way, the semantics consistency and the difference of the sentence representations can be better balanced.

5 EXPERIMENT

5.1 EXPERIMENT SETUP

Following previous works Kim et al. (2021); Gao et al. (2021), we conduct experiments on seven standard STS tasks and take the results as the main comparison of sentence embedding methods. Besides, we also evaluate our approach on seven transfer tasks. For all these tasks, we use the SentEval toolkit Conneau & Kiela (2018) for evaluation.

Semantic Textual Similarity Task. We evaluate methods on 7 STS tasks: STS 2012–2016 Agirre et al. (2016; 2012; 2013; 2014; 2015) , STS Benchmark Cer et al. (2017) and SICK-Relatedness Marelli et al. (2014). These datasets contain pairs of two sentences, whose similarity scores are labeled from 0 to 5. The relevance between gold annotations and the scores predicted by sentence representations are measured by the Spearman correlation. Following the suggestions from previous works Gao et al. (2021); Yan et al. (2021), we directly compute the cosine similarity between sentence embeddings for all STS tasks.

Transfer Tasks. We also evaluate methods on the following transfer tasks: MR Pang & Lee (2005), CR Hu & Liu (2004), SUBJ Pang & Lee (2004), MPQA Wiebe et al. (2005), SST-2 Socher et al. (2013), TREC Voorhees & Tice (2000) and MRPC Dolan & Brockett (2005). Following previous works Gao et al. (2021), we incorporate a logistic regression classifier on top of (frozen) sentence representations for different tasks, and follow default configurations from SentEval.

Baseline Methods. We compare L2P-CSR to the following sentence representation methods:

(1) **GloVe** Pennington et al. (2014): It averages GloVe embeddings of words as the representation.

(2) **USE** Cer et al. (2018): It utilizes the Transformer model to encode sentences and learns the objective of reconstructing the surrounding sentences within a passage.

(3) **CLS** and **Mean** Devlin et al. (2019): These methods adopt the `[CLS]` embedding and mean pooling of token representations as sentence representations, respectively.

Table 1: Sentence embedding performance on STS tasks (Spearman's correlation). The best performance methods in each group are denoted in bold. †: results from Kim et al. (2021); ‡: results from Gao et al. (2021); all other results are reproduced or reevaluated by ourselves.

	Models	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
Non-BERT	GloVe (avg.)†	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
	USE†	64.49	67.80	64.61	76.83	73.18	74.92	76.69	71.22
BERT-base	CLS†	21.54	32.11	21.28	37.89	44.24	20.30	42.42	31.40
	Mean†	30.87	59.89	47.73	60.29	63.73	47.29	58.22	52.57
	+Flow‡	58.40	67.10	60.85	75.16	71.22	68.66	64.47	66.55
	+Whitening‡	57.83	66.90	60.90	75.08	71.31	68.24	63.73	66.28
	+Contrastive(BT)†	54.26	64.03	54.28	68.19	67.50	63.27	66.91	62.63
	+SG-OPT†	66.84	80.13	71.23	81.56	77.17	77.23	68.16	74.62
	+SimCSE	68.40	82.41	74.38	80.91	78.56	76.85	72.23	76.25
	+Ours	**70.21**	**83.25**	**75.42**	**82.34**	**78.75**	**77.80**	**72.65**	**77.20**
	+Prompt	71.56	84.58	76.98	**84.47**	**80.60**	81.60	69.87	78.54
	+Prompt+Ours	**72.34**	**84.81**	**78.13**	84.16	80.58	**82.04**	**71.13**	**79.03**
BERT-large	CLS†	27.44	30.76	22.59	29.98	42.74	26.75	43.44	31.96
	Mean†	27.67	55.79	44.49	51.67	61.88	47.00	53.85	48.91
	+Flow†	62.82	71.24	65.39	78.98	73.23	72.72	63.77	70.07
	+Whitening	64.34	74.60	69.64	74.68	75.90	72.48	60.8	70.35
	+Contrastive(BT)†	52.04	62.59	54.25	71.07	66.71	63.84	66.53	62.43
	+SG-OPT†	67.02	79.42	70.38	81.72	76.35	76.16	70.20	74.46
	+SimCSE	70.88	84.16	76.43	84.50	79.76	79.26	73.88	78.41
	+Ours	**71.44**	**85.09**	**76.88**	**84.71**	**80.00**	**79.75**	**74.55**	**78.92**
	+Prompt	**73.29**	86.39	77.90	85.18	79.97	81.92	71.26	79.42
	+Prompt+Ours	73.14	**86.78**	**78.67**	**85.77**	**80.32**	**82.23**	**72.57**	**79.93**
RoBERTa-base	CLS†	16.67	45.57	30.36	55.08	56.98	45.41	61.89	44.57
	Mean†	32.11	56.33	45.22	61.34	61.98	54.53	62.03	53.36
	+Whitening‡	46.99	63.24	57.23	71.36	68.99	61.36	62.91	61.73
	+Contrastive(BT)†	62.34	78.60	68.65	79.31	77.49	79.93	71.97	74.04
	+SG-OPT†	62.57	78.96	69.24	79.99	77.17	77.60	68.42	73.42
	+SimCSE	70.16	81.77	73.24	81.36	80.65	80.22	68.56	76.57
	+Ours	**71.69**	**82.43**	**74.55**	**82.15**	**81.81**	**81.36**	**70.22**	**77.74**
	+Prompt	73.94	**84.74**	77.28	**84.99**	81.74	81.88	69.50	79.15
	+Prompt+Ours	**74.97**	83.63	**78.28**	84.86	**82.03**	**82.77**	**71.26**	**79.69**
RoBERTa-large	CLS†	19.25	22.97	14.93	33.41	38.01	12.52	40.63	25.96
	Mean†	33.63	57.22	45.67	63.00	61.18	47.07	58.38	52.31
	+Whitening	64.17	73.92	71.06	76.40	74.87	71.68	58.49	70.08
	+Contrastive(BT)†	57.60	72.14	62.25	71.49	71.75	77.05	67.83	68.59
	+SG-OPT†	64.29	76.36	68.48	80.10	76.60	78.14	67.97	73.13
	+SimCSE	72.86	83.99	75.62	84.77	81.80	81.98	71.26	78.90
	+Ours	**73.29**	**84.08**	**76.65**	**85.47**	**82.70**	**82.15**	**72.36**	**79.53**
	+Prompt	73.24	83.08	77.97	84.03	81.57	82.85	73.28	79.43
	+Prompt+Ours	**73.65**	**84.08**	**78.29**	**85.36**	**82.15**	**83.70**	**73.47**	**80.10**

(4) **Flow** Li et al. (2020): It is a flow-based model that applies mean pooling on the layer representations and maps the outputs to the Gaussian space as sentence representations.

(5) **Whitening** Huang et al. (2021):It uses the whitening operation to refine representations and reduce dimensionality.

(6) **Contrastive (BT)** Fang & Xie (2020): It utilizes back-translation as data augmentation for the contrastive learning of sentence representations.

(7) **SG-OPT** Kim et al. (2021): It proposes a contrastive learning method with a self-guidance mechanism to improve sentence embeddings.

(8) **SimCSE** Gao et al. (2021): It propose a simple contrastive learning framework that utilizes dropout as perturbation to make data augmentation.

(9) **Prompt** Jiang et al. (2022): It devises a manual template as the prompt to reduce the token embedding bias, and uses the `[MASK]` embedding as the sentence representation. As it only revises the input format, our proposed L2P-CSR can be combined with it to improve the perturbations.

Table 2: Transfer task results of sentence embedding models (Accuracy). We highlight the highest performance among models with the same base model.

	Model	MR	CR	SUBJ	MPQA	SST	TREC	MRPC	Avg.
BERT-base	CLS‡	78.68	84.85	94.21	88.23	84.13	91.40	71.13	84.66
	Mean‡	78.66	86.25	94.37	88.66	84.40	**92.80**	69.54	84.94
	+SimCSE‡	80.41	85.30	94.46	88.43	85.39	87.60	71.13	84.67
	+Ours	**82.07**	**87.74**	**94.95**	**89.54**	**87.84**	85.18	**75.34**	**86.09**
BERT-large	CLS	74.67	85.77	93.26	80.61	85.44	65.26	69.14	79.16
	Mean	78.55	**89.97**	**96.20**	87.21	88.07	85.53	74.43	85.71
	+SimCSE	84.48	88.95	95.43	88.10	88.88	83.82	74.58	86.32
	+Ours	**88.86**	89.54	95.45	**89.47**	**90.14**	**85.68**	**76.47**	**87.94**
RoBERTa-base	CLS	58.09	69.09	84.60	71.69	79.47	31.64	70.71	66.47
	Mean	75.67	**88.78**	**96.16**	86.12	87.50	83.05	73.63	84.42
	+SimCSE‡	79.67	84.61	91.68	85.96	84.73	**84.20**	64.93	82.25
	+Ours	**79.67**	88.30	94.27	**87.70**	**87.50**	81.14	**76.47**	**85.01**
RoBERTa-large	CLS	58.21	69.38	93.25	75.13	83.03	56.91	70.76	72.38
	Mean	68.19	**89.81**	**96.92**	88.82	**89.33**	86.94	74.90	84.99
	+SimCSE‡	**80.83**	85.30	91.68	86.10	85.06	**89.20**	75.65	84.83
	+Ours	80.12	88.53	94.07	**88.92**	87.04	83.05	**76.84**	**85.51**

5.2 Main Results

To verify the effectiveness of L2P-CSR, we select BERT/RoBERTa-base/large as the backbones.

Semantic Textual Similarity. Table 1 shows the results of different methods on 7 STS tasks. Based on the results, we can find that non-BERT methods mostly perform better than native PLM representations based baselines (*i.e.,* CLS and Mean). The reason is that the PLM native representations are easy to suffer from the anisotropy problem. For other PLM-based approaches, first, we can see that Flow and Whitening achieve similar results and outperform the native PLM representations based baselines by a margin. The reason is that the two methods adopt additional post-processing strategies. Second, contrastive learning based methods mostly outperform other baselines. The reason is that contrastive learning can improve both the alignment between semantically related positive pairs and the uniformity of the whole representation space, which is helpful to improve sentence representations. Furthermore, SimCSE performs better than most baselines. The reason is that the dropout perturbation rarely changes the semantics of the sentence representations.

In addition, we can see that L2P-CSR performs better than all baselines without prompt. As a comparison, our L2P-CRS incorporates a safer perturbation mechanism that only weakens the influence of tokens and features on the sentence representations, and devises a learning algorithm to adapt the perturbation to the dynamically updated sentence representations during training. In this way, the safer mechanism guarantees the semantic consistency, and the learning algorithm helps augment high-quality examples to guide the sentence representation learning.

Finally, we can see that contrastive learning with prompt Jiang et al. (2022) achieves the best performance among all the baselines. The reason is that the manual prompt is able to reduce the token embedding bias. By combining prompt and our L2P-CSR, we can see that the performance of sentence representations can be further improved. It indicates that our approach is a general framework and can effectively improve the perturbations of contrastive learning approaches.

Transfer Tasks. For Transfer tasks, we select commonly-used CLS and Mean as baselines. We also select SimCSE since it performs well in STS tasks. We show the results in Table 2. Among all the PLM-based methods, the performance order is mostly consistent across all datasets, *i.e.,* CLS $<$ SimCSE $<$ Mean $<$ Ours. First, SimCSE performs not well. It indicates that the dropout perturbation may be not suitable to learn representation for Transfer tasks. Second, we can see that Mean performs well in several tasks. The reason may be that mean pooling the token representations can capture token-level characteristics. Finally, L2P-CSR performs better than all baselines on average. The reason is that our proposed learnable perturbation strategy makes the adaptive perturbations better guide the contrastive learning of sentence representations.

Table 3: Ablation and variation studies of our approach.

Model	STS-B	SICK-R
BERT-base+Ours	**77.80**	72.65
w/o feature mask	70.91	**73.01**
w/o token mask	74.05	70.71
w/o learning to perturb	75.09	70.34
BERT-base+Emb Noise	75.95	70.73
BERT-base+Continual Mask	75.05	70.54

Table 4: Results on a subset of tasks from GLUE.

Models	WNLI	QNLI	QQP	SST-2	STS-B
BERT_{base}	33.80	91.29	91.08	92.20	89.71
+Ours	**36.62**	**91.56**	**91.14**	**93.35**	**89.86**

Comparing the performance of different base models, we can find that large models always perform better than their base versions. It indicates that more parameters lead to better representations. Besides, BERT-based approaches perform better than RoBERTa-based ones. The reason is that RoBERTa removes the next sentence prediction task, which can capture sentence-level semantics.

5.3 Further Analysis

Ablation Study. Our proposed L2P-CSR designs two weakening masks and a learning algorithm for perturbation updates. To verify the effectiveness of these modules, we conduct an ablation study in Table 3. First, we see that removing each module would lead to a performance drop. It indicates that all these modules are important. Besides, removing the feature mask results in a larger performance drop. One possible reason is that the feature dimensions are always high in PLMs, perturbing features can better guide the contrastive learning. Emb Noise and Continual Mask are the variations of our framework. Emb Noise follows the VAT paradigm Zhu et al. (2020) that adds gradient-based noise on the embedding layer, and the Continual Mask directly utilizes the probability vectors to weaken the representations. As shown in Table 3, they perform no better than our approach. The reason is that they perturb more radically on the sentence representations, which may hurt their semantics.

GLUE Experiments. In this part, we consider evaluating if our L2P-CSR could enhance the sentence representations when fine-tuning on downstream tasks. Therefore, we evaluate our framework using BERT-base model on a subset of tasks from GLUE Wang et al. (2018), *i.e.,* WNLI, QNLI, QQP, SST-2, and STS-B. Instead of adding a logistic regression classifier on freezing sentence representations, we fine-tune the learned model parameters by our approach on these tasks. Results are shown in Table 4. First, we can find that our framework improves the performance of BERT-base in these tasks. It indicates that our framework enhances sentence representations rather than hurting them. Second, it can be found that our method brings about a larger improvement on the WNLI and SST-2 tasks. It shows that our approach is particularly helpful for these tasks.

6 Conclusion

In this paper, we propose a framework L2P-CSR that learns to perturb for contrastive learning of unsupervised sentence representations. In this framework, we incorporate a safer perturbation mechanism that only weakens the representations from the perspectives of token and feature, and design a gradient-based algorithm to optimize the perturbation to adapt itself to the dynamically changing sentence representations during training. Our approach makes the perturbation able to generate informative perturbations especially for the intermediate sentence representation, and meanwhile guarantees semantic consistency. Experimental results have shown that our approach outperforms several competitive baselines.

REFERENCES

Eneko Agirre, Daniel M. Cer, Mona T. Diab, and Aitor Gonzalez-Agirre. Semeval-2012 task 6: A pilot on semantic textual similarity. In *NAACL-HLT*, pp. 385–393, 2012. URL https://aclanthology.org/S12-1051/.

Eneko Agirre, Daniel M. Cer, Mona T. Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. *sem 2013 shared task: Semantic textual similarity. In **SEM*, pp. 32–43, 2013. URL https://aclanthology.org/S13-1004/.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel M. Cer, Mona T. Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. Semeval-2014 task 10: Multilingual semantic textual similarity. In *COLING*, pp. 81–91, 2014. doi: 10.3115/v1/s14-2010. URL https://doi.org/10.3115/v1/s14-2010.

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. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *NAACL-HLT*, pp. 252–263, 2015. doi: 10.18653/v1/s15-2045. URL https://doi.org/10.18653/v1/s15-2045.

Eneko Agirre, Carmen Banea, Daniel M. Cer, Mona T. Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *COLING*, pp. 497–511, 2016. doi: 10.18653/v1/s16-1081. URL https://doi.org/10.18653/v1/s16-1081.

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. Universal sentence encoder for english. In *EMNLP*, pp. 169–174, 2018. doi: 10.18653/v1/d18-2029. URL https://doi.org/10.18653/v1/d18-2029.

Daniel M. Cer, Mona T. Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *ACL*, pp. 1–14, 2017. doi: 10.18653/v1/S17-2001. URL https://doi.org/10.18653/v1/S17-2001.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, volume 119, pp. 1597–1607, 2020. URL http://proceedings.mlr.press/v119/chen20j.html.

Alexis Conneau and Douwe Kiela. Senteval: An evaluation toolkit for universal sentence representations. In *LREC*, 2018. URL http://www.lrec-conf.org/proceedings/lrec2018/summaries/757.html.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pp. 4171–4186, 2019. doi: 10.18653/v1/n19-1423. URL https://doi.org/10.18653/v1/n19-1423.

William B. Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *IWP@IJCNLP*, 2005. URL https://aclanthology.org/I05-5002/.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and GPT-2 embeddings. In *EMNLP-IJCNLP*, pp. 55–65, 2019. doi: 10.18653/v1/D19-1006. URL https://doi.org/10.18653/v1/D19-1006.

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

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *CoRR*, abs/2104.08821, 2021. URL https://arxiv.org/abs/2104.08821.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 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)*, pp. 879–895, 2021.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. doi: 10.1109/CVPR42600.2020.00975. URL https://doi.org/10.1109/CVPR42600.2020.00975.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. Learning distributed representations of sentences from unlabelled data. In *NAACL-HLT*, 2016. doi: 10.18653/v1/n16-1162. URL https://doi.org/10.18653/v1/n16-1162.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *ICML*, 2019.

Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *SIGKDD*, 2004. doi: 10.1145/1014052.1014073. URL https://doi.org/10.1145/1014052.1014073.

Junjie Huang, Duyu Tang, Wanjun Zhong, Shuai Lu, Linjun Shou, Ming Gong, Daxin Jiang, and Nan Duan. Whiteningbert: An easy unsupervised sentence embedding approach. *CoRR*, abs/2104.01767, 2021. URL https://arxiv.org/abs/2104.01767.

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

Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. Self-guided contrastive learning for BERT sentence representations. In *ACL*, pp. 2528–2540, 2021. doi: 10.18653/v1/2021.acl-long.197. URL https://doi.org/10.18653/v1/2021.acl-long.197.

Quoc V. Le and Tomás Mikolov. Distributed representations of sentences and documents. In *ICML*, volume 32, pp. 1188–1196, 2014. URL http://proceedings.mlr.press/v32/le14.html.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. On the sentence embeddings from pre-trained language models. In *EMNLP*, 2020. doi: 10.18653/v1/2020.emnlp-main.733. URL https://doi.org/10.18653/v1/2020.emnlp-main.733.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. Fast, effective, and self-supervised: Transforming masked language models into universal lexical and sentence encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1442–1459, 2021.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. A SICK cure for the evaluation of compositional distributional semantic models. In *LREC*, 2014. URL http://www.lrec-conf.org/proceedings/lrec2014/summaries/363.html.

Takeru Miyato, Andrew M. Dai, and Ian J. Goodfellow. Adversarial training methods for semi-supervised text classification. In *ICLR*, 2017. URL https://openreview.net/forum?id=r1X3g2_xl.

Jianmo Ni, Gustavo Hernández Ábrego, Noah Constant, Ji Ma, Keith B Hall, Daniel Cer, and Yinfei Yang. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. *arXiv preprint arXiv:2108.08877*, 2021.

Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *ACL*, 2004. doi: 10.3115/1218955.1218990. URL https://aclanthology.org/P04-1035/.

Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL*, pp. 115–124, 2005. doi: 10.3115/1219840.1219855. URL https://aclanthology.org/P05-1015/.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, pp. 1532–1543, 2014. doi: 10.3115/v1/d14-1162. URL https://doi.org/10.3115/v1/d14-1162.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, pp. 1631–1642, 2013. URL https://aclanthology.org/D13-1170/.

Ellen M. Voorhees and Dawn M. Tice. Building a question answering test collection. In *SIGIR*, pp. 200–207, 2000. doi: 10.1145/345508.345577. URL https://doi.org/10.1145/345508.345577.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *EMNLP 2018*, pp. 353, 2018.

Kexin Wang, Nils Reimers, and Iryna Gurevych. Tsdae: Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 671–688, 2021.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Lang. Resour. Evaluation*, 2005. doi: 10.1007/s10579-005-7880-9. URL https://doi.org/10.1007/s10579-005-7880-9.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL-HLT*, pp. 1112–1122, 2018. doi: 10.18653/v1/n18-1101. URL https://doi.org/10.18653/v1/n18-1101.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. CLEAR: contrastive learning for sentence representation. *CoRR*, abs/2012.15466, 2020. URL https://arxiv.org/abs/2012.15466.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. Consert: A contrastive framework for self-supervised sentence representation transfer. In *ACL/IJCNLP*, 2021. doi: 10.18653/v1/2021.acl-long.393. URL https://doi.org/10.18653/v1/2021.acl-long.393.

Dejiao Zhang, Shang-Wen Li, Wei Xiao, Henghui Zhu, Ramesh Nallapati, Andrew O Arnold, and Bing Xiang. Pairwise supervised contrastive learning of sentence representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5786–5798, 2021a.

Dejiao Zhang, Wei Xiao, Henghui Zhu, Xiaofei Ma, and Andrew O Arnold. Virtual augmentation supported contrastive learning of sentence representations. *arXiv preprint arXiv:2110.08552*, 2021b.

Yanzhao Zhang, Richong Zhang, Samuel Mensah, Xudong Liu, and Yongyi Mao. Unsupervised sentence representation via contrastive learning with mixing negatives. In *AAAI*, 2022.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. Freelb: Enhanced adversarial training for natural language understanding. In *ICLR*, 2020. URL https://openreview.net/forum?id=BygzbyHFvB.

Algorithm 1: The algorithm of our L2P-CSR framework.

Input: The unlabeled dataset $\mathcal{X} = \{x_i\}_{i=1}^{n}$, Pre-training language model $f(\cdot)$, Training epoch number m, Gradient-based update steps for perturbations t, Perturbed layer number k,

Output: Fine-tuned PLM $f(\cdot)$.

Initialize $f(\cdot)$ from pre-trained checkpoint.
for $i = 1, \ldots, m$ **do**
  for *batch* $\in \mathcal{X}$ **do**
    Initialize the weakening probability vectors $\mathcal{P}_{tok}^{(l)}$ and $\mathcal{P}_{fea}^{(l)}$ using Eq. 4
    // Iteratively update the perturbations via contrastive objective maximization
    for $j = 1, \ldots, t$ **do**
      Generate the token-level and feature-level weakening masks $\mathbf{M}_t^{(l)}$ and $\mathbf{M}_f^{(l)}$ using Eq. 1
      Perturb the representations of the first k-th layer using Eq. 2 and Eq. 3
      Generate perturbed sentence representations $\tilde{\mathbf{h}}_i$ and $\tilde{\mathbf{h}}_i^+$ by PLM $f(\cdot)$
      Calculate the gradient of weakening masks $\mathbf{M}_t^{(l)}$ and $\mathbf{M}_f^{(l)}$ using Eq. 7 and Eq. 8
      Update the weakening probability vectors $\mathcal{P}_{tok}^{(l)}$ and $\mathcal{P}_{fea}^{(l)}$ using Eq. 5 and Eq. 6
    end
    // Using the perturbations to augment positive examples for contrastive learning
    Generate the token-level and feature-level weakening masks $\mathbf{M}_t^{(l)}$ and $\mathbf{M}_f^{(l)}$ using Eq. 1
    Perturb the representations of the first k-th PLM layer using Eq. 2 and Eq. 3
    Generate perturbed sentence representations $\tilde{\mathbf{h}}_i$ and $\tilde{\mathbf{h}}_i^+$ by PLM $f(\cdot)$
    Update the sentence representations by SGD using contrastive learning objective Eq 9
  end
end

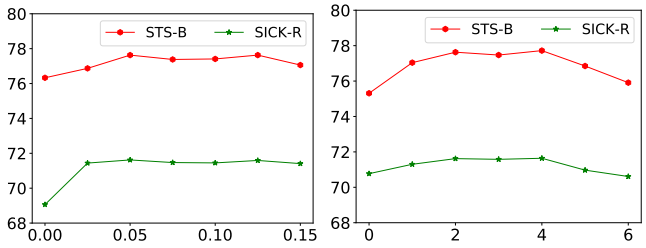
A Algorithm

We present the overall algorithm in Algorithm 1, which depicts the detailed process of our framework L2P-CSR. We first initialize PLM $f(\cdot)$ from its pre-trained checkpoint. Then, for each batch of data, we adopt a learnable perturbation for contrastive learning. Concretely, we first initialize the token-level and feature-level weakening probability vector $\mathcal{P}_{tok}^{(l)}$ and $\mathcal{P}_{fea}^{(l)}$ using Eq. 4. Then, we iteratively update the perturbation via contrastive objective maximization. At each iteration, we first generate the token-level weakening masks $\mathbf{M}_t^{(l)}$ and feature-level weakening masks $\mathbf{M}_f^{(l)}$ to perturb the representations of the first k-th layer of the PLM, and then calculate the gradients of weakening masks $\mathbf{M}_t^{(l)}$ and $\mathbf{M}_f^{(l)}$ to update the weakening probability vectors using Eq. 5 and Eq. 6. After t-turn iterations, we can obtain the optimized weakening probability vectors to generate the most confusing perturbed representations $\tilde{\mathbf{h}}_i$ and $\tilde{\mathbf{h}}_i^+$. Finally, we perform contrastive learning using Eq 9 to update the sentence representations. Specifically, we adopt $\tilde{\mathbf{h}}_i$ and $\tilde{\mathbf{h}}_i^+$ as positive examples, and utilize stochastic gradient descent (SGD) to optimize the parameters of the sentence encoder (*i.e.,* PLM).

B Implementation Details

We implement our L2P-CSR based on Huggingface's transformers packages [2]. We start from pre-trained checkpoints of BERT Devlin et al. (2019) or RoBERTa Liu et al. (2019), and add an MLP layer on top of the [CLS] representation as the sentence embedding. Following SimCSE Gao et al. (2021), we use 100,000 sentences randomly drawn from English Wikipedia as the training corpus. During training, we train our models for 1 epoch with a batch size of 512 and temperature $\tau = 0.05$ using the AdamW optimizer. The learning rate of the model parameters is set as 1e-5 for RoBERTa-base and 3e-5 for other models. The mask threshold ϕ is set as 0.05 and we weaken the embedding layer and the first two layers. The learning rate of the weakening probability vectors is 0.5. We keep the

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



(a) Masking Threshold ϕ (b) Weakened Layer Number k

Figure 2: Performance comparison w.r.t. ϕ and k.

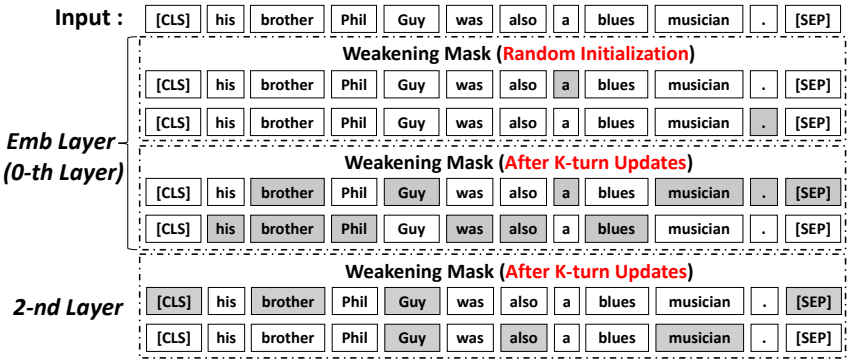


Figure 3: Visualization of token masks. Gray denotes the representation of the token is weakened (*i.e.,* halved).

default dropout layer in PLMs. We evaluate the model every 150 training steps on the development set of STS-B and keep the best checkpoint for the final evaluation on test sets.

C Hyper-parameters Analysis

For hyper-parameter analysis, we study the masking threshold ϕ that is used to produce the masks, and the Weakened layer number k. Both hyper-parameters are important in our framework. Concretely, we evaluate our model with varying values of ϕ and k on the STS-B and SICK-R tasks using BERT-base model. The results are shown in Figure. 2.

In Figure. 2a, we can see that too large or too small ϕ leads to a performance drop. One possible reason is that larger ϕ may hurt the semantics of the representations and smaller ϕ cannot lead to effective perturbations. In Figure. 2b, we can find that the best performance is achieved when masks are applied on the first two or four layers. Since perturbing these layers are more suitable for contrastive learning.

D Case Study

In this part, we show the case study about the learnable perturbations in Fig 3. First, we can see that after K-turn updates, the weakening masks can cover most keywords, *e.g.,* brother and Phil. And the masks for the two views are usually complementary since they weaken different keywords. It indicates that our approach can encourage the perturbation to capture key semantics and augment differential views. This characteristic is essential for effective contrastive learning. Besides, we can see that in the lower layers (0-th), there are more weakened tokens than the higher layer (2-nd). The reason may be that perturbation on lower layers is easier to generate differential representations.